DIVERGENT MICROBIAL PROFILES IN TUMOR AND ADJACENT NORMAL TISSUE ACROSS CANCER TYPES

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

BIOMEDICAL SCIENCES (CLINICAL RESEARCH)

BY

REBECCA M RODRIGUEZ

Dissertation Committee:

Brenda Y Hernandez (Chairperson)
Youping Deng (Co-Chairperson)
Vedbar S Khadka
Amy Brown
Scott Kuwada
Loic LeMarchand
Lynne Wilkens, Outside Member

Purposely Left Blank

DEDICATION

To the patients who volunteered sample tissues from which the data was derived;
To all my family and friends who have lost their battle, and those who continue to battle with all forms of cancer

ACKNOWLEDGEMENTS

Many thanks to Dr. Brenda Y. Hernandez, Dr. Youping Deng and Dr. Vedbar S. Khadka. Without their mentorship, restless support and encouragement this work would not have been possible. My sincere gratitude to Dr. Amy Brown for her continuous advice and editorial contributions; to Dr. Lynne Wilkens, Dr. Scott Kuwada, Dr. Loïc LeMarchand and Dr. John Chen for their guidance, flexibility and input to this work. Many thanks to the Cancer Center Pathology Shared Resources for their work with the population blocks in the validation phase of this project. My deepest appreciation to Dr. Mark Menor from the JABSOM Bioinformatics Core, as his dedicated support ensured the completion of this project.

My heartfelt thanks to all those who encouraged and supported me throughout this journey. I am especially grateful to my dear husband, Anibal and my children, Anibal and Axel, who sacrificed so much so that I could achieve this dream.

The results published here are in part based upon data generated by the Cancer Genome Atlas managed by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). Information about TCGA can be found at http://cancergenome.nih.gov. All human data were handled in accordance with TCGA Data Use Certification Agreement and Data Access Request (DAR) 57292; Project-14778 (Deng, PI), Request ID 57292-2 and 57292-4 (renewal 10/10/2018) for level-1, controlled-data access phs000178 versions v9.p8 and v10.p8.

This research was funded by the National Institute of Minority Health and Health Disparities, and the Health And Wellness Achieved by Impacting Inequalities (OlaHAWAII) pilot grant number 2U54MD007601-32 (Khadka, PI). The Bioinformatics Core is supported by National Institute of Health (NIH) Grant numbers 5P30GM114737, P20GM103466 and U54MD007584. R. M. Rodriguez was supported in part by the Population Sciences in the Pacific Epidemiology Program at the University of Hawaii Cancer Center.

The Hawaii Tumor Registry is part of the National Cancer Institute's (NCI) Surveillance, Epidemiology, and End Results (SEER) Program and operate affiliated Residual Tissue Repository (RTR). Tissue blocks from the Hawaii RTR were used for PCR validation. Exempt IRB approval was obtained prior to any study procedures relating to use of tissue blocks (protocol number 2018-00178).

STRUCTURE OF THE DISSERTATION

This dissertation is structured using UHM manuscript option. The research proposal is introduced first in Chapter I, followed by review of the literature in Chapter II. The research proposal was successfully submitted to OlaHAWAII Team Pilot Grant for funding. A portion of the literature review was submitted for publication as a review article to a peer review journal. A series of manuscripts on the microbial associations with cancer pathogenesis are presented in Chapter III. Subsequently, a summary of the results, overall conclusion and future research are discussed in Chapters IV. Work completed during the life of the project and supplemental tables or figures are included as appendices in Chapter V.

PROJECT SUMMARY

Background: There is growing evidence that microbial variation can influence cancer development, progression, response to therapy, and outcomes. We wanted to examine the microbial composition of paired tumor and adjacent normal tissue across various cancer types in order to provide an improved understanding of microbial diversity and abundance patterns of the tumor microenvironment and their influence on clinical presentation and survival. **Methods**: Using raw whole exome sequencing data from 22 cancer types from the The Cancer Genome Atlas (TCGA) network, we examined differential relative abundance and diversity data in primary tumor and adjacent solid tissue normal (adjacent normal), nine of which are presented in this work. Data were processed through a bioinformatics pipeline designed to extract microbial profiles from human sequencing data based on PathoScope 2.0. Differential abundance and diversity metrics were calculated using R-package tools to compare primary tumor and adjacent normal within cancer types and across cancer cohorts correlating to clinical features including histologic and pathologic features and survival data. Analyses were controlled for demographic, exposures and batch effects. Findings were then validated by qPCR for selected cancer types with tissue from the Hawaii Tumor Registry RTR. **Results**: As part of a pilot project we have created microbial composition and diversity profiles for a subset of solid tumors within TCGA cancers building a platform for ongoing and future studies. We screened over 10,000 files encompassing a total of 1,838 paired cases from which 813 are discussed here. From those, 767 tumors and 753 adjacent normal samples had positive microbial (viral and bacterial) sequence reads detection. Microbial composition and diversity (richness, within sample alpha diversity, evenness and beta diversity) varied across cohorts with similar patterns at the phylum level in compositional structures. Bacterial shifts were evident in tumor compared to adjacent normal. Proteobacteria phyla was observed to be increased in tumors of all cohorts except for STAD, where Proteobacteria species were reduced and Firmicutes levels increased. Differences between patient samples were evident at the higher taxonomic levels. Differential abundance analyses revealed significant differences in stomach adenocarcinoma (STAD) and colon adenocarcinoma (COAD). Compared to adjacent normal, tumor samples were found to have lower number of species present overall and lower diversity indices. We found significant association between microbial relative abundance and diversity to clinicopathological presentation and survival dependent on race in some cancers, particularly those of infectious origin like STAD and LIHC. **Conclusion**. This project demonstrates the feasibility of the utilization of exome sequencing data to derive complex microbial data with easy to interpret results. This project facilitates the understanding of the role of bacteria play in cancer pathogenesis across different race groups as demonstrated in LIHC and STAD cancer cohorts. In these cancers, relative abundance was associated with tumor stage and overall survival days and within sample diversity was associated with race in fully adjusted models.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

TABLE OF ACRONYMS

CHAPTER I. RESEARCH PROPOSAL

1.1 **Objectives**

The goal of this research was to profile the microbial composition of human tumors and to evaluate their role in disease progression and survival. Comparisons were made by race and other characteristics. Utilizing a comprehensive approach, we compared human derived microbial data between primary tumor, henceforth “tumor” and adjacent solid tissue normal, henceforth “adjacent normal” sequencing samples from The Cancer Genome Atlas (TCGA) cancer cohorts.

1.2 **Specific Aims**

1.2.1 Primary Aims

- Determine the microbial relative abundance in tumor and adjacent normal samples across cancer types using sequencing data and bioinformatics tools.
- To determine cancer associations by correlating abundance and diversity metrics to clinical features and survival data.

1.2.2 Assumptions

- Infected tissue contains both human and microbial nucleic acids
- Pathogen-derived sequences can be detected after subtraction of human sequences (Weber et al. 2002, Xu et al. 2003, Kostic et al. 2011)

1.2.3 Hypotheses

- Microbial relative abundance in cancer tissue can be derived from human whole exome sequencing data similar to that derived from whole transcriptome and whole genome sequencing methods.
- Consistent with the literature, it was further hypothesized that

  (a) Adjacent tissue would be associated with greater microbial species diversity
  and
  (b) That location of the colonization and patterns of co-occurrence would be associated with cancer status and therapeutic effect (determined by survival) with varying patterns by race and ethnicity.

This is because microbial composition between and within each body organ is distinct which can help drive functional inter-individual variations and determinants of disease (Schwabe and Jobin 2013). Prevailing infectious disease exposures differences, and burden of infection-associated cancers vary among racial and ethnic groups. Taking these together, microbial abundance and diversity patterns could contribute to racial disparities.

1.2.4 Secondary aims

- Experimentally validate bioinformatics microbial composition findings from TCGA cohort against formalin-fixed paraffin-embedded (FFPE) tissue blocks

1.3 **Significance**

Identification of co-factors involved in widening racial-related cancer disparities is essential to precise diagnostic and treatment strategies. It is accepted that microbiota can influence carcinogenesis and tumor progression through inflammatory and immune response (Zitvogel et al. 2017). Recent studies point out substantial differences in the gut microbiota composition in healthy individuals and cancer patients from different racial and ethnic groups (Brooks et al. 2018, Farhana et al. 2018). Bacteria and bacterial metabolites interactions within the tumor microenvironment may have important diagnostic and therapeutic implications in the reduction and elimination or racial differences. An important component of this project was to evaluate perturbations in microbial diversity of the tumor tissue across different cancer types while correlating to clinical features and overall survival. This study provides much-needed information on differential bacterial associations with cancer, especially given their potential use in microbe-based prevention strategies and possible new targeted therapies.

1.4 **Experimental Approach**

Using TCGA data in a Pan-cancer approach, we proposed to retrospectively examine the relationship between bacteria and cancer pathogenesis to cancer survival among racial and ethnic groups. To do this, we derived bacterial and viral species relative abundances from human whole exome sequencing data using a bioinformatics pipeline designed to generate microbial profiles from binary version of Sequence Alignment/Map (BAM) files from solid tissue TCGA cancer cohorts. We used relative abundance for each microbe and diversity metrics to correlate clinical features including basic demographics (age at diagnosis, sex, vital status, race and ethnicity), exposures (alcohol and smoke), tumor stage, tumor grade, treatment type (whenever available), histopathology and survival (days to death and vital status). This will help determine if a relationship exists between bacteria and cancer pathogenesis. Findings from bioinformatics filtering of TCGA data were validated in a cross-sectional fashion using FFPE tissue from a racially diverse sample from the Hawaii Tumor Registry Residual Tumor Repository (Hawaii's RTR). We performed quantitative-polymerase chain reaction (qPCR) using commercially available primers and de-

identified archival tissue from selected cancers. Authorization for use of archival tissue was obtained from UH IRB (Exempt Protocol 2018-00174).

1.5 **Research Design**

This study used secondary data analyses and cross-sectional design to retrospectively examine the relationship between bacteria and cancer pathogenesis.

1.5.1 Study population

This project is part of an OlaHAWAII Team Pilot Grant (2U54MD007601-32). The data used in this study was derived from The Cancer Genome Atlas (TCGA) consortium (phs000178 versions v9.p8 and v10.p8, downloaded October 2017 through November 2018). TCGA is a harmonized cancer data infrastructure that is continuously being updated. As of June 2017 (data release 7.0) there were 274,724 files constituting 11,160 unique cases across 33 cancer types including bladder carcinoma (BLCA), breast carcinoma (BRCA), cervical squamous carcinoma (CESC), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), cancers of the head & neck (HNSC), kidney chromophobe (KICH), kidney renal cell (KIRC), kidney papillary carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell (LUSC), ovarian carcinoma (OV), pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), rectal adenocarcinoma (READ), sarcoma (SARC), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA), thyoma (THYM) and uterine corpus and endometrial carcinoma (UCEC). TCGA is a collaborative effort between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) which effectively combine sample and clinical data from multiple platforms with the goal of developing better strategies for diagnosing, treating and preventing cancer. Datasets are available in different access tiers, open-access (level-3 including de-identified clinical data and demographics, gene expression, copy number alterations, epigenetic data, compiled data summaries, and anonymized amplicon DNA data) and controlled-access (levels 1 and 2 including single nucleotide, variant call, microsatellite instability, whole exome, whole genome, tumor miRNA, tumor mRNA and raw data) (http://tcga-data-nih.gov). Although not part of TCGA primary goals, these efforts and availability of cross-platform data enable detection of microbial commonalities in tumor types and examination of bacterial effects on cancer modulation processes across different sample populations. TCGA samples originate from a diverse populations across the United States, while the majority of samples were from Whites (73%, regardless of Hispanic origin), there is a representative sample from major racial and ethnic groups (**Table 1**). We requested access to level 1, raw whole exome sequencing data and corresponding de-identified clinical and overall survival data (Project # 14778). We anticipate there are approximately 8,000 solid tumor and adjacent normal raw sequencing files with available clinical data that can be examined for bacterial composition (**Table 2**). In a comprehensive approach we expect to compare the microbial (bacterial and viral) profiles of available solid tissue cancers contained within the TCGA consortium and correlate differences and commonalities

with clinical data meeting selection criteria. For this study, 22 solid tumors were selected on the basis of whole exome sequencing data availability for tumor and adjacent normal specimens with at least 13 total samples. We elected to examine solid tumors to directly extract microbial content related to tumor microenvironment and to reduce introduction of bias by examining blood normal derived sequences. All sequencing case pairs meeting selection criteria were downloaded and screened form microbial content. Cases were defined as solid tumor cancer types within TCGA that had human aligned sequencing reads, paired primary tumor and adjacent normal raw exome sequences in BAM file format, plus available clinical data. Paired cases were selected at 1:1 ratio for the bioinformatics interrogation. Each of the selected cancer types had a minimum of 13 samples.

Table 1 *Racial and ethnic distribution in TCGA cases*

Race or ethnic group	Cases	%
White, Caucasian	8186	73.4
Black or African American	934	8.4
Asian	675	6.0
Native Hawaiian, Other Pacific Islander	13	0.1
Native American, Alaskan Native	27	0.2
Not Reported	1325	11.9
Not Hispanic or Latino	8173	63.9
Hispanic or Latino	377	2.9
Not reported	2610	20.3

Table displays racial and ethnic group breakdown in TCGA datasets as of data release 7.0 (June 2017).

A separate IRB Protocol submission was made for the validation of findings and use of archival tissue from the Hawaii's RTR. The RTR and Hawaii's Tumor Registry are part of the NIH, Surveillance Epidemiology and End Results Program (SEER). RTR comprises of a uniquely racially diverse collection of FFPE tumor specimens from cancer patients diagnosed within the catchment area of Hawaii's Tumor Registry. Validation with archival tissue from a racially diverse population is vital to the understanding of racial and ethnic variation in cancer incidence and survival. Archival tissue was requested for selected cancer types for which microbial profiles showed significant results. Tissue was requested at similar 1:1 ratio tumor and adjacent normal.

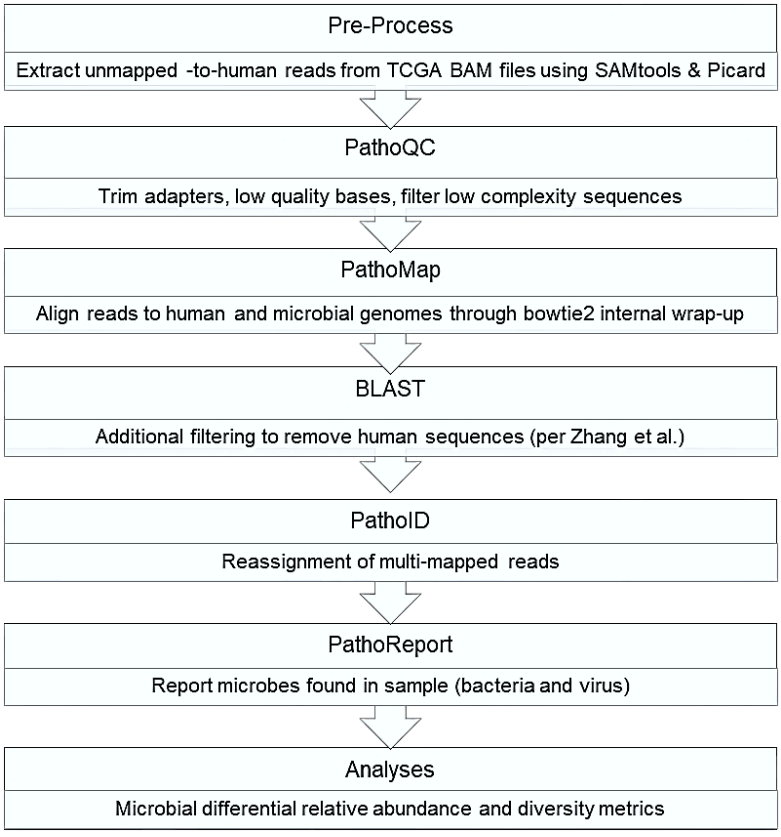
1.5.2 Methods and Planned Statistical Analyses

Specific Aim 1. To determine the microbial relative abundance in tumors and adjacent normal paired samples across cancer types using exome sequencing data and bioinformatics tools

We determined microbial relative abundance from unmapped to human reads using PathoScope 2.0 and R shiny app package PathoStat (Hong et al. 2014, Manimaran S 2017) (**Figure 1**). PathoScope is a well

described bioinformatics tool developed to detect microbial reads present in clinical or environmental samples' sequencing data (Hong et al. 2014). PathoStat is designed to work with PathoScope generated reports (Manimaran S 2017). In brief, selected sequencing reads were quality filtered, trimmed and compiled for data analysis using SAMtools and Picard. Additional filtering was completed by computational subtraction against human reference genome (hg38) and then aligned to known microbial genomes (custom library). Relative abundance was calculated for each microbe (taxa) based on normalized values in tumor and normal adjacent tissue for all cancer types that met described case selection criteria. Viral DNA presence, mainly HPV, HBV and EBV, was used as internal pipeline validation and for microbial co-occurrence analyses. Our pipeline was designed to identify DNA sequences as such, RNA viruses like HCV were not detected. Diversity measurements, alpha (diversity within each sample) and beta (differences in composition among samples) per cancer type and across cancers were calculated.

Figure 1 *Bioinformatics Pipeline*



Bioinformatics pipeline designed to extract microbial profiles from human sequencing data using modified PathoScope methods and additional filtering steps described by Zhang et al. [19]. Viral sequences detection used as internal validation and co-occurrence examination.

Shannon Diversity Index (alpha within sample diversity) was used to compare between tumor and its paired adjacent normal. We analyze difference of proportions of bacterial presence within cancer types by racial/ethnic groups comparing tumor to its non-tumor normal. We will use contingency tables and non-parametric statistical testing such as Chi-Square. Differential microbial abundance between tumor and its paired adjacent normal in each cancer type was determined using edgeR (Robinson, McCarthy, and Smyth 2010). Bacterial taxa with false discovery rate (FDR) adjusted p value < 0.05 were considered significantly different at both genus and species level. Major bacteria taxa present in 50% or more of the cases (per cancer type) was defined as core microbiome. Cancer types found to have significant microbial read differences were designated for experimental validation using archival (formalin-fixed, paraffin-embedded, or FFPE) tumor tissue. All microbial data derived from human DNA sequences were kept in an electronic database as described in the Data Management Plan (Appendix B. **Data Management**123).

Table 2 *Available paired cases by cancer cohort in The Cancer Genome Atlas*

TCGA Cohort	Total Files	Total Cases	Total Paired
BLCA	455	412	37
BRCA	1202	1050	137
CESC	313	305	8
CHOL	71	51	20
COAD	561	443	88
ESCA	248	184	64
HNSC	587	527	69
KICH	123	66	57
KIRC	652	345	250
KIRP	381	290	88
LIHC	460	376	84
LUAD	853	582	200
LUSC	836	502	221
OV	634	460	96
PAAD	221	185	36
PRAD	608	498	106
READ	182	168	18
SARC	277	255	22
STAD	530	443	88
THCA	598	502	98
THYM	136	123	13
UCEC	599	553	38
TOTAL	**10,527**	**8,320**	**1,838**

Available WXS primary tumor and solid tissue normal sample sequence files from TCGA solid tumor cancer cohorts with at least 10 paired cases from which microbial data could be derived. Data release 10.1

Specific Aim 2. To determine cancer association by correlating microbial abundance & diversity data to clinical features & survival data.

We used relative abundance for each microbe and its correlation to clinical features including basic demographics (age at diagnosis, sex, race and ethnicity), tumor stage, tumor grade, treatment type (whenever available), histopathology, exposure (alcohol and smoke when available) and survival (days to

death and vital status) to derive cancer associations. To determine the association between differences in relative abundance in tumor and its adjacent normal and clinical features, paired or unpaired t-test and analysis of variance (ANOVA) were used for two- and multi-group comparisons, respectively. Chi-square test were used for categorical data. Using differentially abundant microbes and significant clinical features as predictors, Cox proportional hazards regression analyses were performed to evaluate the associations between each of the potential biomarkers and overall survival outcomes per cancer types.

Specific Aim 3. Experimental validation of bioinformatics microbial findings from the TCGA cohort using FFPE tissue from cancer registry-based population

TCGA microbial associations for selected cancer types of the stomach and lung were validated in a cross-sectional fashion with archival population blocks. One hundred-twenty (120) tissue blocks per cancer were requested at 1:1 ratio per paired-case (tumor and its paired adjacent normal) in representative distributions for Hawaii's largest racial and ethnic groups. The Cancer Center Pathology Share Resource completed tissue retrieval, cut & slide, sectioning, pathology review and nucleic acid extractions. DNA was extracted using appropriate purification kit to maximize DNA yield from FFPE (Carrick et al. 2015). PCR was performed for species-specific bacterial DNA presence using commercial probe/primers qPCR kits (Qiagen Microbial DNA qPCR kit, Qiagen, USA). Samples were run in duplicate and included positive and negative controls with each run. Bacterial detection results were correlated with de-identified demographic, clinical and survival data provided by Hawaii's RTR. Experimental data was compared to bioinformatics results. Due to budget constraint, we validated up to six (6) of the most abundant or differentially abundant bacteria per cancer type. It was determined that to attain enough power, 30 samples with expected discordant proportion ratio of 0.5 and odd ratios of 0.18 would have at least 80% power to detect the bacterial association per cancer type at the significance level of 0.05 using McNemar's test.

1.5.3 Strengths of the research proposal

The proposed study is unique in being the first to evaluate cross-cancer associations between bacterial abundance and cancer pathogenesis as determined by survival and clinical data with bioinformatics and experimental validation design within the local Hawaii population. The proposed study also utilizes matched read pairs strengthening the results of the microbial commonalities and differences. The availability of TCGA data makes it possible to accomplish the study aims in a timely and cost effective manner. Hawaii's distinctive racially diverse population makes it possible to extrapolate the findings to the general population.

1.5.4 Limitations and Alternative Strategies

Our sample size is limited by the number of quality paired sequencing files available per cancer type in the TCGA databases and budget constraints for validation. Our ability to establish causality is limited by the cross-sectional study design. We are limited to pre-existing data which was not collected at time of

pathogen exposure or prior to cancer diagnosis. Also, our pipeline was designed to detect DNA species, as such one of our limitations is not being able to detect RNA viruses.

1.5.5 Changes to planned analyses

All analyses in this research were performed using R version 3.5.1 (2018-07-02) and MS Excel (v. 2013). Multiple problems were encountered while processing data through the R shiny app PathoStat since the study's initial pipeline validation with SRA files and were unable to complete differential relative abundance analyses using PathoStat. Alternative methods within R-packages were utilized similar to those used by PathoStat to determine relative abundance and diversity metrics including vegan R-package (version 2.5-3), microbiome R-package (version 1.3.3), and phyloseq (version 1.25.3). A list of R tools utilized is found in Appendix D2, pp142.

Due to the nature of the study data and the relative low counts of microbial reads generated from exome sequencing data compared to whole genomic and transcriptomic methods we were unable to consistently use edgeR (Robinson, McCarthy, and Smyth 2010) or similar methods such as DESeq-2 (Love, Huber, and Anders 2014)for differential analyses with count data for all cancer types. Differential relative abundance was analyzed using Wilcoxon-Signed Rank and Rank Sum tests using normalized proportion data.

TCGA bioinformatics interrogation was shortened due to data volume and time constraints. In addition, archival tissue was limited due to budgetary restrictions and avoiding depletion of sample pool. As such convenience sampling was utilized for selection of cases for validation portion of this study.

1.6 **Impact**

Profiling the microbial composition of human tumors is a daunting task that is being facilitated by new sequencing technology, bioinformatics tools and merging cooperative networks. The TCGA, one of such networks, has catalogued over 11,000 cases across 33 tumor-types. The Pan-Cancer analysis project is an effort to compare tumor omics characteristics and clinical data across major tumor types from data drawn from the TCGA network. In a Pan like approach using whole exome sequencing and clinical data from the TCGA consortium this research examined the relationship between bacteria and cancer pathogenesis. This provided a cross-cancer view of tumor-bacterial associations, enabling the confirmation or rejection of established or suspected associations, discovery of new associations, identification of patterns of co-occurrence, and examination of host-interaction effects.

1.7 **Literature cited**

Located in Appendix **E. Literature Cited, complete list**, pp.147

CHAPTER II. REVIEW OF THE LITERATURE

The literature review is divided into two sections, *"**The role of tumor microbiota in cancer pathogenesis and the impact on racial related disparities,**"* and *"**Identification of racial-related microbial differences across cancers from human-derived sequencing data.**"* Portions of the literature review were submitted for publication to a peer review journal.

2.1 **The Role of Tumor Microbiota in Cancer Pathogenesis and the Impact on Racial-Related Disparities**

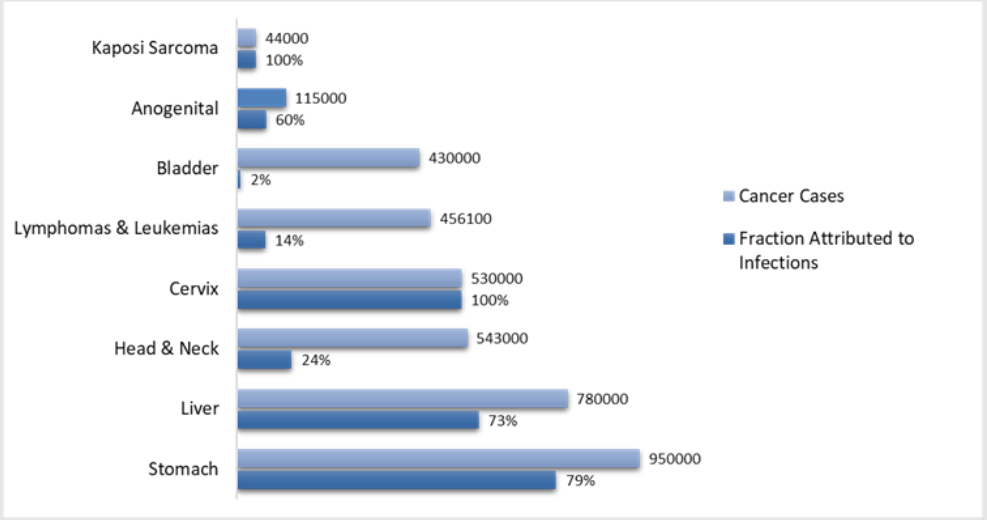
2.1.1 **Abstract**

Cancer is the second leading cause of death in the United States. Etiology varies by cancer type with observed differences among racial and ethnic groups in incidence and mortality rates. It is understood that while cancer affects all persons, some ethnic and racial groups are at higher risk of developing, suffering complications, and dying from cancer. In the United States, despite reduced incidence and mortality rates, racial and ethnic minorities continue to show striking differences in cancer outcomes. There is evidence that infection burden may contribute to health disparities. Recent studies elude to distinct tumor microbial patterns contribution to variation observed among several cancer types, which warrants further exploration to understand roles of tumor microbiota behind racial disparities and to provide new diagnostic and therapeutic strategies. The review covers what is known about the divergent relationship of tumor microbiota in cancer pathogenesis and their potential impact on racial related disparities.

2.1.2 **Introduction**

Cancer is one of the leading causes of death worldwide. In United States, 1.7 million individuals are diagnosed and nearly 600,000 die from the disease each year (Siegel, Miller, and Jemal 2015, 2016, 2018). Worldwide, one in five cancer cases are attributed to infectious agents (Plummer et al. 2016, de Martel et al. 2012, Parkin 2006). **Figure 2**. In the United States out of the top 10 leading causes of incidence and cancers related deaths, 9 cancer types are known to have marked racial differences including cancers of the lung, prostate, breast, colorectal, pancreas, liver, kidney, stomach and cervical (Hernandez and Goodman 2004, Gourin and Podolsky 2006, Curtis et al. 2008, Singh 2012, Li et al. 2013, Li et al. 2014, Sturtz et al. 2014, White et al. 2014, Daraei and Moore 2015, NCI 2015, DeSantis et al. 2016, Ha et al. 2016, Ragin et al. 2016, Setiawan et al. 2016, Miller et al. 2017, Scelo et al. 2017, Zhang et al. 2017). Of these, 4% of cases are thought to be infection-associated, primarily cases of the liver, stomach and cervical cancers (IARC 2012, Parkin 2006, de Martel et al. 2012). A considerable proportion of these cancer cases remain with unknown etiological factors underscoring the importance of understanding the interplay of contributory factors to include microbial contributions.

Figure 2 *Global Distribution of cancers attributable to infectious agents*



Bar graph displaying the infection attributable fraction of cancers new cancer cases worldwide. Light bar represents new cancer cases, dark bar represents attributable fraction in percent of total cases. The remaining fraction of new cases have considerable proportion of unknown etiology. Data source: Plummer et al., Globocan and the American Cancer Society (https://www.cancer.org/cancer/cancer-causes/).

2.1.3 **Microbiota in Cancer Pathogenesis**

The human microbiome is defined as the aggregation of microorganisms including viruses, bacteria, and eukaryotes that live in and on our bodies which genomes contribute to our broader genetic portrait (Human Microbiome Project 2012, Rogers 2016, Ursell et al. 2012). The microbiota within each organ system is distinct which can drive functionally relevant inter-individual variations and determinants of disease (Schwabe and Jobin 2013, Bik et al. 2010). Bacterial community variations, the production of bacterial metabolites, and microbial interaction with the human host have been attributed to detrimental and beneficial tumoral effects since the 18th century (Nauts 1982, Nauts 1989). Persistent and chronic exposure to infectious agents may initiate or promote cancer (Parkin 2006, Plummer et al. 2016). Equally, any agent that stimulates immune defenses can minimize the incidence and be beneficial once cancer develops (Blaser 2008, van Tong et al. 2017, Parsonnet 1995, de Martel et al. 2005, Chang and Parsonnet 2010). This highlights the unique agonistic and antagonistic effects of the human microbiome in cancer progression and has become an area of intense exploration.

The microbiological infections worldwide contribution to cancer pathogenesis ranges from 4% to 31% (Plummer et al. 2016, Kuper, Adami, and Trichopoulos 2000, Cogliano et al. 2011, WHO 2018). While contribution by specific viral pathogens is firmly established, for bacteria, archea, and eukaryotes agents remains controversial particularly for bacterial members of the commensal microbiota. Here we briefly discuss known and suspected oncogenic viral and bacterial agents, archea and eukaryotes, which are less common, are not discussed.

2.1.3.1 Infectious etiological factors

2.1.3.1.1 Viral

Viruses are responsible for at least 10% of all human cancers (Moore and Chang 2010) and constitute about 63.5% of the new cancers attributed to infections (Plummer et al. 2016). Viruses share common characteristics in the development of cancer and have been reviewed in detail by several researchers (Parkin 2006, Kuper, Adami, and Trichopoulos 2000, De Flora and La Maestra 2015, Burnett-Hartman, Newcomb, and Potter 2008). There are six generally accepted viral carcinogens: human herpes virus 4 also known as Epstein Barr virus (HHV-4 or EBV), hepatitis B virus (HBV), hepatitis C virus (HCV), human papilloma virus (HPV), human T-cell virus-1 (HTLV-1), and human herpes virus 8 also known as Kaposi Sarcoma Herpes virus (HHV-8 or KSV). The seventh viral etiology, human immunodeficiency virus-1 (HIV-1), is included because it indirectly promotes carcinogenesis caused by other viruses through immunosuppression (Beuth 2005, Monographs 2012).

Other viruses have been investigated with conflicting or limited evidence regarding their carcinogenic effects including polyomaviruses. There is recent limited evidence which proposes a causal relationship between Merkel cell polyomavirus (MCPyV) in Merkel cell carcinoma (Cogliano et al. 2011, Shuda et al. 2008) whereas conflicting evidence remains for the human polyomavirus JC as a risk factor in colorectal cancer (Laghi et al. 1999, Lundstig et al. 2007). Recent studies have found associations between viral presence with distinct gene expression patterns and cancer aggressiveness (Cancer Genome Atlas Research 2014, Khoury et al. 2013). For example, although controversial, presence of HHV-4 exists in aggressive breast tumors (Mazouni et al. 2011).

2.1.3.1.2 Bacterial

Bacteria have long been associated with cancer, either promoting or protecting depending on site and cell type of the colonization (Schwabe and Jobin 2013, Chang and Parsonnet 2010, Mager 2006). High site-specific colonization and microbial-host dynamics permits a dual role in cancer pathogenesis. Inflammation microenvironments created by the host microbiota can promote accumulation of mutations and epigenetic changes that lead to cell modulating and tumor-promoting effects (Hattori and Ushijima 2016, Elinav et al. 2013). Conversely, bacteria and its products can prevent, arrest or regress cancer progression by acting directly or indirectly on host inflammatory response and cellular pathways (Nair, Kasai, and Seno 2014). This is exemplified by Helicobacter pylori contrasting role in gastric adenocarcinoma (increased risk) and esophageal adenocarcinoma (decreased risk). In a study by de Martel and colleagues, *Helicobacter pylori* colonization was associated with 37% risk reduction of esophageal adenocarcinoma development in persons 50 years and younger regardless of smoking status (de Martel et al. 2005). *Helicobacter pylori* is widely accepted to play a role in carcinogenesis. In fact according to IARC, is the only bacterial agent recognized to have a definite role in the development of various cancers including gastric adenocarcinoma and MALT gastric lymphoma, while suspected to play a significant role in other malignancies (Parsonnet et al. 1993, Wotherspoon et al. 1991, Murphy et al.

2014, Pellicano et al. 2004, Hong et al. 2012). Like *Helicobacter pylori*, *Salmonella enterica* serovar Typhi is believed to contribute to the development of gallbladder cancer in endemic regions and hepatocellular carcinoma in chronic carriers (Koshiol et al. 2016, Caygill et al. 1994). However, it is believed to have beneficial effects in the treatment of cancers like melanoma (Hayashi et al. 2009). Other bacteria species have been found associated with cancer modulation. *Streptococcus bovis* (Gold, Bayar, and Salem 2004) *Bacteroides fragilis* (Thomas et al. 2016) and more recently *Fusobacterium nucleatum* for example have been found to be enriched in colorectal tumors by several teams (Kostic et al. 2013, Marchesi et al. 2011, Castellarin et al. 2012, Warren et al. 2013, Kumar et al. 2016). *Fusobacterium nucleatum* has varying degrees and distinct patterns of microbial co-occurrence which can result in carcinogenic and anti-carcinogenic effects (Zitvogel et al. 2017, Marchesi et al. 2011). Effects which are dependent on host-microbial interactions and location of colonization (Chang and Parsonnet 2010).

Figure 3 *Proposed mechanisms by which bacteria contribute to the alterations and the carcinogenic process*

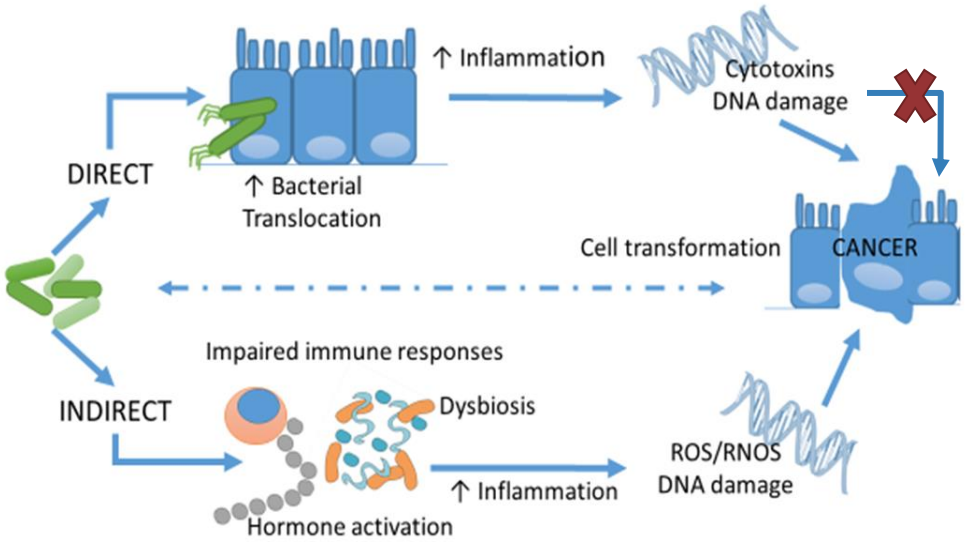


Figure displays proposed bacteria direct and indirect mechanisms of action in infection and inflammation associated cancers. Bacteria may directly transform cells by disruption of membrane structures that lead to an increase in bacterial translocation. This in turn can result in establishment of chronic inflammation and double strand DNA damage by bacterial cytotoxins. Bacterial cytotoxins may also lead to beneficial effects by mode of cytotoxins initiated apoptosis, halting the progression of the disease. Bacteria can also indirectly alter hormone metabolism, increase immune altered responses. Altered patterns leading to dysbiosis can in turn cause chronic inflammation and the release of reactive oxygen (ROS) and nitrogen species (RNOS) induced tissue damage and cellular transformation. Alterations of microbial community could result in beneficial or detrimental effects to the tumor microenvironment.

2.1.3.1.3 Mechanisms of action

It is understood that exposure to microbial agents alone is not sufficient to induce the carcinogenic process (Moore and Chang 2010). Rather it is the interaction of multiple host risk factors and molecular events induced by microbial agents and their products that lead to transformed cancer cells. Common

mechanisms of action include direct expression of viral oncogenes or direct and indirect alteration of the inflammatory response, as is in the case of bacterial agents (Moore and Chang 2010, Elinav et al. 2013). The mechanisms by which viral agents contribute to pathogenesis have been reviewed in detail (Elinav et al. 2013, Moore and Chang 2010, Burnett-Hartman, Newcomb, and Potter 2008, Hattori and Ushijima 2016, De Flora and Bonanni 2011, Kuper, Adami, and Trichopoulos 2000) and are not covered here. Mechanisms by which bacteria contribute to the alterations and the carcinogenic process are not well understood. It is known however, that similar to viral mechanisms, persistent and chronic infections may initiate the process or promote an established cancer (Nauts 1989, Monographs 2012). Alteration of the microbial community could also result in beneficial effects to the tumor microenvironment. In fact, according with the literature, any agent capable of stimulating host immune defenses can minimize the incidence and be beneficial to established tumors. Current proposed mechanisms of bacteria modulation have been summarized into 3 general pathways that can be protective or carcinogenic through direct or indirect, chronic inflammation or immunosuppression (Monographs 2012). These involve actions that alter microbial-host dynamics, disrupt cellular activities or alter the immune cascade in response to the bacterial agent, bacterial toxic metabolites, or bile acid degradation **Figure 3** (Elinav et al. 2013, Schwabe and Jobin 2013, Blaser 2008, Chang and Parsonnet 2010, Mager 2006, Parsonnet 1995, Gagnaire et al. 2017).

Modification of the immune cascade in response to infection or dysbiosis is one of the most important aspects of tumor-microenvironment cross talk (Elinav et al. 2013). Altered host-dynamics can increase bacterial translocation as a direct consequence to changes in microbial composition resulting in increased inflammation. This interaction can result in tumor suppression activities via activation of cancer-preventing phytochemicals. Bacterial products and bacterial metabolites may have protective effects on survival, reduced growth of cancer cells, or modulate anticancer immunosurveillance at local or distant sites (Zitvogel et al. 2017). Butyrate for example, which has anti-inflammatory properties is thought to be protective while secondary bile acids are thought to be carcinogenic (Parsonnet 1995, Bordonaro, Lazarova, and Sartorelli 2008). Notably African Americans, who suffer disproportionately higher incidence and mortality rates for many cancers, have been found to have lower levels of butyrate and butyrate producing microbes than other racial or ethnic groups (Hester et al. 2015). These variations in microbial composition may be in part responsible for the carcinogenic process in susceptible or genetically predisposed populations and can assist in understanding of patient inter-variability of racial related cancer disparities (Moore and Moore 1995, Goyal et al. 2016). More recently, it was demonstrated that pro-inflammatory bacteria were overexpressed in African Americans compared to Caucasian Americans colorectal cancer patients (Farhana et al. 2018). In the same study, Caucasian Americans were found to have higher diversity and greater proportion of probiotic species (Farhana et al. 2018). This study provided evidence of potential microbial contribution to colorectal cancer racial-related cancer disparities.

New microbial contributions to cancer, whether beneficial or detrimental, are being discovered by improved techniques and integrated data networks and have become the focus of multiple studies (Sobhani et al. 2011, Thomas et al. 2016, Marchesi et al. 2011, Warren et al. 2013, Kumar et al. 2016, Riley et al. 2013, Chan et al. 2016, Cavarretta et al. 2017, Thompson et al. 2017, Xuan et al. 2014, Yow et al. 2017, Gopalakrishnan, Spencer, et al. 2018). Several studies have found that specific bacterial taxa are consistently identified in cancer tissue either at increased levels such as *Fusobacteria*, *Alistipes*, *Porphyromonadaceae*, *Coriobacteridae*, *Staphylococcaceae*, *Akkermansia* and *Methanobacteriales*, or at decreased levels like *Bifidobacterium*, *Lactobacillus*, *Ruminococcus*, *Faecalibacterium*, *Roseburia*, and Treponema compared to adjacent or control tissue (Sun and Kato 2016). Differential abundance patterns, provide great potential to examine proposed direct and indirect mechanisms of action as well as their contributions to cancer pathogenises.

2.1.3.1.4 Co-occurrence and aggressiveness

Changes in species diversity or abundance may contribute to increased cancer incidence in susceptible populations. Competitive interactions of microbial agents are more apparent at broader taxonomic levels. These interactions provide information at the population level for each cancer type allowing for racial and ethnic differences to be evaluated. At the species level, patterns of co-occurrence provide individual patient information allowing for personalized therapies (Bik et al. 2010). Taxonomic level analyses of the microbiome have revealed several microbial candidates implicated in pathology of disease (Xie et al. 2016, Kumar et al. 2016). Xie et al. (2016) recently showed that co-abundance patterns in the gut alter liver disease progression including liver cancer; while Kumar identified possible target proteins in *Fusobacterium nucleatum* and host interaction that explain colorectal cancer progression as a result of co-infection. These findings can be applied to preventive or complementary therapies. Viral-bacterial co-occurrence has been identified to modulate tumor aggressiveness. Based on epidemiological and geographic correlations it is suggested that viral agents may interact with bacteria resulting in more aggressive tumors. For example, it is recognized that HHV-4 infected stomach tumors are molecularly distinct. HHV-4 is thought to interact with *Helicobacter pylori* however insufficient evidence exists. In hepatocellular carcinoma co-infection with HBV with HCV and their interaction between proteins HBx, HCV core and NS5a can also lead to more aggressive tumors. In addition other exposures can act as co-factors altering the tumor microenvironment and disease outcomes such as alcohol consumption, smoking co-morbidities and betel nut chewing (De Flora and Bonanni 2011, Hernandez et al. 2017). Furthermore, historical epidemiological data suggests an antagonistic relationship between infectious disease prevalence and cancer incidence and cancer specific mortality (Nauts 1989). This antagonistic relationship may be affected by the timing of the infection (acute versus chronic), and the interaction of the pathogen with the human-host (Hoption Cann, van Netten, and van Netten 2006, Cruz-Munoz and Fuentes-Panana 2017). Concurrent acute infections with *Staphylococcal* or *Streptococcal* species are recognized to have a protective effect in times leading to complete tumor regression (Jeljaszewicz,

Pulverer, and Roszkowski 1982). Yet for infections that become chronic, the opposite effect is observed. Endemic prevalence can be protective for certain cancer types while it has a promoting role in others as demonstrated by *Helicobacter pylori* and *Salmonella* serotype infections (Iyer et al. 2016).

2.1.4 **Commensals and Pathobionts**: Cancer Treatment and the Role of Acute Infections

In cancer patients, treatment is known to alter the host immune responses, induce system barrier failures, and promote infections (Zitvogel et al. 2017, Iida et al. 2013, Routy et al. 2018). Bacterial acute infections are a common cause of morbidity and mortality in cancer patients and susceptibility is driven by cellular disturbances caused by the type of treatment or the cancer type and stage (Rolston 2017, Culakova et al. 2014). The underlying causes and the effects during and after cancer treatment are not well studied and much remains unknown. There is however, important temporal variation involving timing of the treatment, number of circulating immune cells (particularly neutrophils) and the characteristics of the bacterial agent with differing patient outcomes (Attie et al. 2014). Attie et al. (2014) recently examined cancer-specific survival correlation to acute bacterial infection in a colorectal cancer cohort. Study results suggested that presence of infection during and up to one year after treatment was associated with poorer prognosis independent of age (Attie et al. 2014). In fact, infection is the cause of death in about half the patients with gastrointestinal tumors and sepsis complications developed in about 45% of these where the tumor type is thought to be the predisposing factor (Rolston 2017). The infectious agents attributed to sepsis complications have been found to be associated with cancer type (Danai et al. 2006). In addition, there are significant racial and gender disparities associated with sepsis incidence and survival rates among cancer patients. For example, in a population based longitudinal study (1979 to 2001) risk for sepsis among male patients was approximately 30% higher than that of females, while risk for sepsis for African Americans and other races was twice that of whites (Danai et al. 2006). This is in part due to differences in the microbial gene expressions of each individual affecting differential susceptibility to disease along with varying exposures to microbial agents (Blaser 2008, Wallace, Martin, and Ambs 2011). Indeed acute infections can propel production of exotoxins influencing the course of an established cancer (Attie et al. 2014). It has been demonstrated that cancer patients concurrently infected with Staphylococcal or Streptococcal species may experience spontaneous tumor regression in all cancer types (Nauts 1982). Streptococcal infection have also been identified to contribute to cancer development through chronic infection and inflammation (Biarc et al. 2004). This highlights the importance of co-factors and conditions at initial colonization which may determine pattern of the host-agent interaction (Kuper, Adami, and Trichopoulos 2000). Commensals and pathobionts balance disruptions that can result in non-transient changes to the tumor and adjacent tissue microenvironment which dictate the success of the cancer treatment in certain individuals (Iida et al. 2013, Gopalakrishnan, Helmink, et al. 2018). As such, commensals and pathobionts that exert effects on the tumor and adjacent tissue microenvironment could be the key to overcoming tumor-induced immune tolerance (Garcia-Castillo et al. 2016). Commensal and

pathobionts may play a role in cancer risk as part of our overall genetic makeup (Dethlefsen, McFall-Ngai, and Relman 2007) and may be an important predictor of survival having both anti-tumoral and tumor-tolerant properties (Blaser 2008, Lai et al. 2014, Zitvogel et al. 2017). Commensal bacteria are involved in metabolic processes, inflammation and immunity, while pathobionts support the immune system by assisting in the maintenance of the epithelial barrier and in the immune cell maturation process (Hornef 2015). These groups consist of mutualistic or pathogenic bacteria that are transient or residential members of the host's microbiome which generally do not affect the healthy host unless incited by dysbiosis or altered immune responses (Contreras et al. 2016, Garcia-Castillo et al. 2016)

Commensal bacteria in animal models of tumor microenvironment reveal that treatment is affected by the tumor microbiota (Iida et al. 2013). Iida et at. (2013) examined tumor microenvironment after treatment and observed that pro-inflammatory genes were decreased in the absence of microbiota and that microbiota modulated genotoxicity of therapeutic compounds independent of immune elicited cell death. This study suggests that commensal bacteria differentially affect the type of inflammatory response to different therapies and highlights the potential to improve cancer treatment by manipulating gut microbiome (Iida et al. 2013). We can conclude that microbial-host dynamics lead to cellular changes, which result in subsequent divergent cancer modulating effects. Because of this dual modulating effect and the high specificity of bacteria to host interactions, identification of bacterial associations are important to cancer prevention, diagnosis and treatment strategies. Especially to elucidate potential mechanisms involved in bacterial divergent roles in cancer pathogenesis and their impact in racial and ethnic-related disparities.

2.1.4.1 **Modulating effects**

A multiplicity of interactions might exist between infection, colonization, and host-microbiome dysbiosis that predispose susceptible individuals to cancer, while others are provided protection from the disease. There is growing evidence suggesting a dual modulating effect of microbial communities in cancer pathogenesis (Zitvogel et al. 2017, Beuth 2005, Nair, Kasai, and Seno 2014, Mager 2006). Is debatable however, whether these disturbances are causative or a consequence of cancer pathogenesis with enough evidence to suggest both (Garcia-Castillo et al. 2016). Selective adherence is demonstrated in the presence of differential microbial composition by anatomical site in tumor compared to non-tumor progenitor cells (Garcia-Castillo et al. 2016).

Butyricicoccus, and lactic acid bacteria (LAB), mainly *Lactobacillus* and *Bifidobacterium*, are believed to benefit the host through improved intestinal immune cell function, anti-inflammation, anti-tumorigenesis, and pathogen exclusion (Sun and Kato 2016, Kamiya et al. 2016). The beneficial effects of gut microbiota on the host are mainly mediated by its metabolites. Short-chain fatty acid (SCFA), including acetate, propionate, and butyrate, are the major end-products of gut bacteria fermentation of dietary fiber (Rios-Covian et al. 2016). In the 19th century, Coley first proposed the idea of bacterial infection for

cancer shrinkage(Jeljaszewicz, Pulverer, and Roszkowski 1982). New evidence suggests that targeting the microbiome can improve therapeutic outcomes of anticancer drugs (Gopalakrishnan, Spencer, et al. 2018). *Bifidobacterium* for example is thought to be associated with antitumor effects where oral administration promotes antitumor immunity and facilitates anti-PD-L1 efficacy a checkpoint blockade (Sivan et al. 2015, Routy et al. 2018). In the other hand production of bacteria lipopolysaccharides (LPS) is linked to be detrimental to the host through dysbiosis of gut microbiota and modification of antigen presentation (Jorgensen et al. 2016). Several studies have examined the role of gut microbiota in gastrointestinal and non-gastrointestinal cancer pathogenesis (Sobhani et al. 2011, Golombos et al. 2018, Gopalakrishnan, Spencer, et al. 2018, Farhana et al. 2018, Lakritz et al. 2015). Few have evaluated the role of the microbiota on the tumor and adjacent tissue (Marchesi et al. 2011, Kostic et al. 2012, Xuan et al. 2014, Cao et al. 2016, Thompson et al. 2017, Wang et al. 2017).

2.1.5 **Cancer Racial Related Disparities**

Cancer disparities are observed when in general incidence and mortality rates are improving, yet lag behind for certain minority groups. Major efforts have been taken to elucidate the factors influencing cancer disparities. The primary goal of these efforts is to develop response strategies that eliminate or reduce the marked differences. Despite declining incidence rates around the globe, cancer remains the leading cause of death among Asian Americans, Native Hawaiians and Hispanics (Siegel et al. 2015, Torre et al. 2016). In addition, in the United States, cancer incidence and mortality rates are disproportionately higher among Non-Hispanic Blacks or African Americans for the four most common cancers (breast, lung, prostate and colorectal) compared to Non-Hispanic Caucasian American (Siegel, Miller, and Jemal 2018). Cancers most commonly associated with infectious agents, like stomach, liver and cervical cancers are disproportionately higher among Asian Americans, Native American /Alaskan Native and Hispanics (Siegel et al. 2015, Torre et al. 2016, DeSantis et al. 2016). Although the factors influencing variation in incidence and mortality rates among racial and ethnic subgroups in the United States remain largely unknown, disparities are explained in part by rates of screening & preventive service utilization, lifestyle & behavioral risk patterns, tumor biology, and differences in exposure to cancer-causing infectious agents (Deshmukh et al. 2017, Wallace, Martin, and Ambs 2011). These factors have been reviewed by other authors and are beyond the scope of this work. Briefly, utilization of primary & secondary prevention strategies has been historically lower among racial and ethnic minorities in the United States perhaps due to cultural variations and limited access to quality care (Kagawa-Singer et al. 2010). Racial minorities groups within the United States are less likely to be screened at early stage and more likely to receive subpar treatment when compared to Caucasian counterparts; even when income and health insurance status are equal (Singh and Jemal 2017). This seems in accordance with the observed cancer screening & preventive service utilization patterns among racial and ethnic minorities but does not entirely explains their higher mortality rates. For example, African American women have lower incidence of breast cancer yet suffer poorer outcomes and higher mortality rates than any other

racial or ethnic group largely due to differences the timing of screening type of treatment, and the response to treatment strategies (Wallace, Martin, and Ambs 2011, Newman 2017, Curtis et al. 2008). Similarly, among Asian American, Native Hawaiian and Asian ethnic subgroups (Chinese, Japanese and Filipino and other Asian and Pacific Island ethnic minorities) living in the United States, colorectal cancer screening rates explain in part incidence variation among the ethnic subgroups, although not necessarily mortality rates for all subgroups (Hernandez and Goodman 2004). Native Hawaiian suffer disproportionate mortality rates indicative of various co-factors resulting in more aggressive tumors and poorer survival (Hawaii Cancer Center 2016). Evidence suggests response to treatment and tumor aggressiveness to be affected by tumor microenvironment differences and microbial load (Iida et al. 2013, Vetizou et al. 2015) which can help unravel the factors involved in racial differences.

Epidemiological data has shown that lifestyle and behavioral risk patterns influence the variation in incidence and mortality observed in racial and ethnic minorities and explain in part racial related disparities (Gourin and Podolsky 2006, Curtis et al. 2008, Kagawa-Singer et al. 2010, Iqbal et al. 2015, Ha et al. 2016). Data from migrating populations, like the Multiethnic Cohort study, has provided evidence regarding inter-ethnic variability of cancer risk due to of lifestyle, behaviors and exposures as exemplified by the rates of cancer of Japanese migrants in Hawaii that resemble those of their host rather than the country of their birth (Kolonel, Altshuler, and Henderson 2004, Setiawan et al. 2016). Lifestyle & behavioral risk patterns such as diet, sex behavior, smoking and alcohol consumption can alter epigenetic expression where exposure and susceptibility by cancer type vary among different ethnic groups. These play an important part in the continuum of carcinogenesis affecting tumor biology, genetic and epigenetic expression patterns and difference of exposure to cancer-causing infectious agents which dominate the cancer burden in racial and ethnic minorities un the United States.

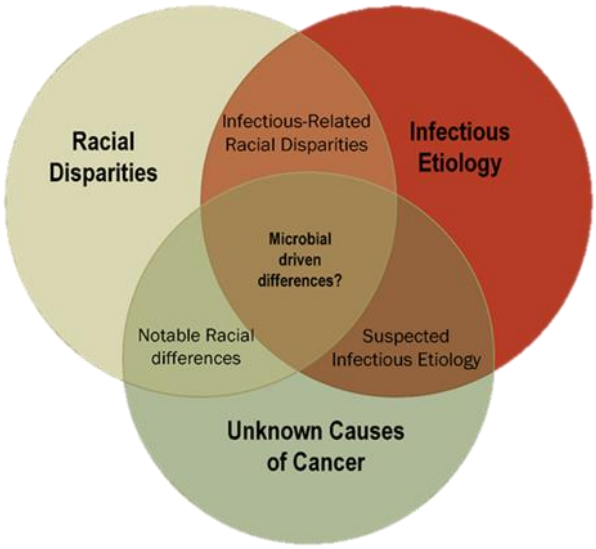
Tumor biology including aggressiveness and differences in exposure to cancer-causing infectious agents are a contributory factors in themselves and explain portion of the persistent racial related cancer disparities in the United States. There is evidence that prevalence of cancer subtypes have distinct and varied patterns within ethnic minorities and likely contribute to more aggressive tumors and poorer outcomes in breast, prostate and colorectal cancers (Newman 2017, Newman and Kaljee 2017, Yamoah et al. 2015, Chen et al. 1997). Exposure to cancer causing agents are also likely to affect tumor biology and is an area of active exploration aside from being a recognizable major risk factor for many of the most common cancers. Yet, differences in overall know risk factors for infection- associated cancers do not explain racial disparities and much is still to be explored (Setiawan et al. 2016). Disparities in these cancer types can be influenced by the prevalence of infectious agents among minority groups and ethnic subgroups (Siegel et al. 2015, Torre et al. 2016). Burden of infection-associated cancers depends on endemic status and access to preventive and therapeutic means as well as the interplay between

behavioral and other environmental factors which predispose racial and ethnic minorities at a greater rate (Setiawan et al. 2016).

2.1.5.1 Impact of the microbiota in racial related disparities

Recent renewed interest has recognized the microbiome as an important contributor in cancer although its role is still unclear. It has been postulated that bacterial community may alter host gene expression that could shed light into racial disparities. For example, Cao et al. (2016) examined viral presence across 23 cancer types and observed that HHV-4 and hepatitis B virus genotypes variants were associated with ethnicity in liver hepatocellular carcinoma and stomach adenocarcinoma among Caucasian and Asian TCGA cohorts (Cao et al. 2016). According to Cao et al. (2016) amino acid distribution of the viral variants separated the Asian ethnic cohorts into distinct groups which provides additional opportunities in the pursuit to understand and eliminate cancer racial disparities and personalized medicine. Understanding the complexity of the interactions between screening utilization, behavioral risk patterns, tumor biology and differences in exposure to cancer-causing infectious agents and how other potential contributory factors, like the very important microbiome, may play an active role in pathogenesis are key to reducing and eliminating racial-cancer disparities. In the United States out of the top leading causes of cancer

Figure 4 *Potential impact of the microbiota in racial related disparities*



Venn diagram showing theoretical intersection between known factors of racial related disparities, infectious etiological factors and unknown causes of cancer. Fraction may be explained by pathogen interactions with host as a potential co-factor in racial related differences.

incidence and mortality, several have both a great proportion of unknown causes and notable racial differences (Siegel, Miller, and Jemal 2018) while others have a great proportion of attributable or suspected microbiological infection etiologies with known racial related differences (Parkin 2006, Plummer et al. 2016). It is not impossible to think that a portion of the unknown causes impacted by racial related differences may be explained by the microbiological attributable causes (**Figure 4**). Several authors have addressed common biological and non-biological contributory factors for the widening racial disparities in cancers of the breast, prostate, lung, stomach, liver, cervix, head & neck and kidney racial disparities. Others have addressed potential microbiological effects in clinical presentation, aggressiveness and outcomes of these cancers. Very few have addressed the potential role of microbiological infections in racial disparities. While contributory factors to racial disparities are beyond the scope of this work, here we briefly discuss cancer specific known racial differences in disease outcomes attributed to a combination of biological and non-biological factors and recent discoveries to further understand potential impact of the microbiota in racial-related differences.

2.1.5.1.1 **Breast**

Breast cancer one of the leading causes of cancer incidence and mortality among females in the United States with 268,600 cases and 41,760 deaths each year (Siegel, Miller, and Jemal 2018). Breast cancer is classified into four classes, Luminal A, Luminal B, HER2-type enrichment, and basal like, with several subtypes based on the estrogen receptor, gene expression and molecular characteristics of the tumor (Cancer Genome Atlas 2012). There is clear evidence of differential distribution of breast cancer classes/subtypes by racial and ethnic groups with impact on racial disparities in disease burden and patient survival. For example, African American and Hispanic females are more likely to be diagnosed with the basal like triple negative (estrogen receptor, progesterone receptor, and human epidermal growth factor negative, TNBC) breast cancer compared to Caucasian American and Asian American females (Howlader et al. 2014, Sturtz et al. 2014). Yet, when considering age and socioeconomic factors, Caucasian American and Hispanic females are at equal risk of being diagnosed with TNBC (Parise and Caggiano 2017). However, African American females have higher probability of dying from breast cancer overall (Siegel, Miller, and Jemal 2018, DeSantis et al. 2016). Compared to Caucasian American, Asian American females have historically lower risk for breast cancer, although true risk may be obscured by data aggregation of ethnic subgroups. In fact, invasive breast cancer is the most commonly diagnosed cancer among Asian American females with Native Hawaiian females' rates similar to Caucasian rates. According to the Multiethnic Cohort Study, piloted by the Universities of Hawaii and California, highest incidence rates are experienced by African American, Japanese, Caucasian American and Native Hawaiian with the latter experiencing the highest rates (Kolonel, Altshuler, and Henderson 2004). Differential distribution of breast cancer classes/subtypes and incidence and mortality patterns among Asians in the United States vary widely across ethnic subgroups. In a study by Telli et al. examining the

population-based distribution of breast cancer subtypes among Asian ethnic groups in California, it was found that four out of six examined ethnic subgroups, Korean, Filipina, Vietnamese and Chinese females were at significantly increased risk of being diagnosed with HER2-positive subtypes compared to Caucasian American females (Telli et al. 2011). These breast cancer subtypes, TNBC and HER2-positive are known to be more aggressive and have poorer survival outcomes among racial minority groups even after controlling for known risk factors (Parise and Caggiano 2017, Telli et al. 2011, Parise et al. 2009). Molecular characterization of breast cancer has discovered distinct genetic and epigenetic abnormalities with distinct clinical manifestations within each subtype including hypomethylation and high expression of DNA repair proteins in basal like breast cancer, high expression of EGFR and HER2 in HER2-type breast cancers, and differential estrogen signaling in RPPA reactive Luminal A and Luminal B breast cancer subtypes (Cancer Genome Atlas 2012) Coincidentally, estrogen metabolism has been associated with tumor microbial diversity (Chan et al. 2016, Xuan et al. 2014). Microbial dysbiosis also affects estrogen levels increasing risk and promoting oncogenesis correlating with gene expression levels, glucoronidation of estrogen, and tumor growth pathways (Chan et al. 2016, Xuan et al. 2014, Thompson et al. 2017). Presence of pathogenic and commensal microorganisms including virus, bacteria and parasites have been found to differentiate TNBC into distinct subtypes (Banerjee et al. 2015, Banerjee et al. 2018)which could assist with diagnostics. Recent studies have found an association between HPV or EBV viral prevalence and tumor progression, aggressiveness and prognosis (Piana et al. 2014, Huo, Zhang, and Yang 2012). Although association remains controversial, positivity could be indicative of ethnic differences. Bacteria are also implicated in the modulation of the host inflammatory response suppressing and/or activating production of pro-inflammatory cytokines and microbial derived signals (Chan et al. 2016, Xuan et al. 2014). Pro-inflammatory cytokines including resistin and IL-6 and other genetic factors have been implicated in African American persistent breast cancer racial disparities (Deshmukh et al. 2015). Microbial community has been observed to produce probiotic effects in certain breast cancer subtypes by enhancing host immune defenses through activation of toll like receptors such as TLR5 by *Sphingobium yanoikuyae* (Chan et al. 2016). Is apparent that microbial community composition and individual species can be beneficial or detrimental factors in pathogenesis hinting at their functional relevance and potential benefit in the understanding of racial related cancer disparities and the pursue to eliminate them.

Banerjee et al. (2015) recently described microbial signatures patterns of co-occurrence in the breast microenvironment among four breast cancer types including triple negative (Banerjee et al. 2015, Banerjee et al. 2018). The study identified two distinct microbial clusters with four subgroups based on viral, bacterial fungal and parasitic signatures in triple negative breast cancer. It is interesting to note that triple negative breast cancer, which is biologically more aggressive than other breast cancer types, shows significant racial disparities for women of West African ancestry (Jiagge et al. 2016, Newman 2017). This

is an important feature that could be indicative of pathogen-host co-evolution and transgenerational inheritance of oncogenic mutations with impact on racial related cancer outcomes.

2.1.5.1.2 **Prostate**

Despite the recent drop in prostate cancer incidence rates it remains the second most common cause of cancer mortality among men (Siegel, Miller, and Jemal 2016, 2018, DeSantis et al. 2016). Prostate cancer accounts for an estimated 164,690 of new cancer cases and 29,430 of cancer deaths in the United States each year (Siegel, Miller, and Jemal 2018). There are approximately seven subtypes defined by specific gene fusions with varied androgen receptor activity and epigenetic profiles (Cancer Genome Atlas Research 2015, Steele et al. 2017). African American males are twice more likely to develop and die from prostate cancer than Caucasian American with poorly understood risk factors (DeSantis et al. 2016). There is conflicting evidence on the association of race and the receipt of treatment, treatment outcomes, overall and disease specific mortality (Shavers and Brown 2002). Prostate cancer specific death disparities have narrowed while overall cancer survival disparities in prostate cancer patients remain (Steele et al. 2017). In a recent cohort study evaluating potential risk factors and clinical relevant variables among 7307 prostate cancer patients and 432 documented prostate cancer specific deaths, Williams et al. (2018) found that after adjusting for treatment and potential attributable risk factors racial disparities persist in prostate cancer specific mortality among African American males (Williams et al. 2018). Even though disparities are palpable in relation to Caucasian American, attributable factors remain elusive thought to be related to unequal treatment, behavioral risk patterns and tumor biology. Recent studies looking at the molecular basis of racial disparities point out to the interaction between genetics and epigenetics in prostate cancer expressed phenotypes and clinical outcomes from androgen biosynthesis and metabolism (Li et al. 2013, Karakas et al. 2017). It has been noted that African American males have higher levels of sex hormones with a dynamic and rapid decline in testosterone levels compared to Caucasian American (Hu et al. 2015, Xu et al. 2015). Interestingly, the microbial community has the ability to alter androgen and estrogen hormone levels directly or indirectly through genetic alterations impacting tumor aggressiveness and has been the focus of multiple efforts (Cavarretta et al. 2017, Yow et al. 2017, Cohen et al. 2005, Sfanos et al. 2008, Yu et al. 2015, Golombos et al. 2018). Characterization of the prostate microenvironment has identified differences in the microbial community composition between prostate tumor tissue and non-tumor specimens, delineating potential interaction and molecular mechanisms that may contribute to disease progression and the observed racial differences in prostate cancer (Cavarretta et al. 2017, Yow et al. 2017, Karakas et al. 2017, Kumar et al. 2018, Porter et al. 2018). These findings collectively can have direct impact in the pursuit to better understand and eliminate disparities not only for incidence, but also in disease outcomes.

2.1.5.1.3 **Lung**

Lung cancer is the number one cause of cancer related death in the United States. In 2018, 25.3% of all cancer deaths were attributed to the disease (Siegel, Miller, and Jemal 2018). Incidence and mortality

rates are higher among males compared to females for all racial and ethnic groups higher among African American compared to non-Hispanic Whites (SEER Cancer statistics 2018). Smoke exposure is the major risk factor for cancers of the lung, yet does not account for all lung cancer deaths and a fraction remains unexplained highlighting the importance of other exposures (CDC 2018). In non-smokers, suggested risk factors involve exposures to chemical and biological agents. Factors affecting racial disparities among lung cancer patients include geographic variations, baseline comorbidities and stage at diagnosis. Likewise these only explain up to 32% of the observed racial differences (Karanth et al. 2018) and much remains unknown. Tumor molecular characteristics have been postulated as possible contributory factors in differences particularly among African American and Caucasian counterparts. A recent study of the molecular landscape of cancer revealed that in non-small cell lung adenocarcinoma, BRAF was mutated at a higher frequency and in squamous cell carcinoma programmed death ligand 1 (PD-L1) expression tended to be lower in African Americans compared to Caucasians (Heath et al. 2018). Microbial dysbiosis may also be correlated with lung cancer development and could explain some of the observed variation. Several studies have linked viral, bacterial and parasitic agents to lung cancer initiation and progression. EBV, HPV and MCV viruses have been associated with lung cancer patients from Asia (De Paoli and Carbone 2013). *Chlamydia pneumoniae* has also been implicated in the development and to contribute to lung metastasis by several studies (Pevsner-Fischer et al. 2016). Coincidentally, Chlamydia pneumoniae seropositivity has been associated with race, where non-White and non-Black lung cancer patients are more likely to have high titers (Mager 2006). However, little is known about the lung microbiome, the evidence is limited and conflicting. Despite limited evidence microbial dysbiosis may be an active participant in lung cancer and perhaps a contributory factor in racial differences with clinical importance that should be further explored.

2.1.5.1.4 **Stomach**

Gastric cancer is the third leading cause of cancer related deaths with approximately 1,033,701 new cancer cases diagnosed throughout the world in 2018 (GLOBOCAN at gco.iarc.fr). Despite continued decrease in incidence and mortality rates, Asian Americans and Native Hawaiians show striking disparities in incidence being twice as likely of being diagnosed with gastric cancer compared to Non-Hispanic whites (Siegel, Miller, and Jemal 2018). Among Hispanic men incidence is 60% higher compared to non-Hispanic whites males while remaining similar to that of other aforementioned minority racial and ethnic groups for Hispanic females (cancer.org). Studies have revealed that differences in tumor biology and infectious etiology contribute to observed differences including *Helicobacter pylori* strain cytotoxin-associated gene A (Cag-A) and vacuolating cytotoxin A (Vac-A) variation and stomach microbiota (Merchant, Li, and Kim 2014, Noto and Peek 2017, Brawner et al. 2017). Gastric cancer and *Helicobacter pylori* contribution to cancer progression is one of the best studied bacterial infection associated cancers, although the actual mechanisms remain largely elusive (Chang and Parsonnet 2010, Parsonnet et al. 1993). Recent advances in sequencing technology have permitted elaboration on the

role of not only *Helicobacter pylori*, but the contribution of gastric microbiota to cancer pathogenesis and their role in the observed differences in different population groups. These have largely concentrated in the gastric environment of cancer patients. Recently, Yu et al evaluated the microbiota of gastric cardia adenocarcinoma tumor and non-tumor tissue paired samples and found that gastric microbiota as a major risk factor *Helicobacter pylori* relative abundance was significantly higher in patients without family history of cancer compared to patients with family history and was associated with tumor grade (Yu et al. 2017). There is also a fair body of literature on the role of HHV-4 on gastric cancers.

2.1.5.1.5 **Colon and rectal**

Despite improved prevention strategies and the recent decrease in mortality, colorectal cancer continues to be the second leading cause of death in both men and women combined in the United States (Siegel, Miller, and Jemal 2018). Incidence and mortality rates vary by racial and ethnic groups with African American and Native American /Alaskan Natives having higher mortality and incidence rates that any other groups (Siegel et al. 2015, Torre et al. 2016, DeSantis et al. 2016). African Americans are 20-30% more likely to be diagnosed with colorectal cancer than their non-Hispanic White counterparts whereas Native American and Alaskan Native groups have marked geographic variations, presumably associated to environmental, cultural and behavioral factors (White et al. 2014). American Asian and Hispanic ethnic subgroups have 10% to 30% lower incidence and mortality rates with great intragroup variability (Siegel et al. 2015, Torre et al. 2016). Although racial disparities have often been attributed to health behaviors and socioeconomic status, the causes for racial disparities in colorectal cancer are of great exploration and debate (DeSantis et al. 2016). Studies suggest that ethnic differences in dietary habits and the interactions with gut microbiota lead to non-transient effects in the intestinal lumen and differential expression of microbial genes or bacterial metabolites in the colon tissue that leads to the accumulation of insults (Kostic et al. 2013, Moore and Moore 1995, Wu et al. 2011). This provides evidence of the potential role of microbes in cancer related differences in this multifactorial disease. Recently, microbial dysbiosis has been identified to play a potential role in racial related differences in African American colorectal cancer patients (Goyal et al. 2016, Farhana et al. 2018). For example, Farhana et al. (2018) examined diversity and abundance of microbial composition in African American and Caucasian American colorectal cancer patients and found distinct bacterial composition among the groups. African American had a greater proportion of unique operational taxonomic units, while Caucasian American had greater diversity. There were significant differences in the ratio of Bacteriodetes to Proteobacteria species at the phylum level in African American patients compared to their Caucasian counterparts.

2.1.5.1.6 **Liver**

Liver cancers are the second most common cause of cancer in the globe. In the United States liver and intrahepatic bile duct cancers are high among Hispanics, Native American and Alaskan Natives, African

Table 3 *Microbes associated with cancer pathogenesis and potential role in racial related disparities*

Cancer type	Microbial Associations	Racial-related differences	Microbiome Studies	References
Breast **TNBC,** **HER2+, and ER+**	Known -none Suspected: -Alistipes spp -HHV-4, HPV, *Bacteroides fragilis,* *Sphingobium yanoikuyae* Microbial dysbiosis	- Black or African American higher risk of TNBC compared to White -Asian American subgroups show significant intra-ethnic variability among Korean, Filipina, Vietnamese and Chinese	Thompson 2017 Chan 2016 Xuan 2014 Banerjee 2015	Howlader 2014 Sturtz 2014 Parise 2017, 2009 Telli 2011 Huo 2012 Deshmukh 2015 CGAN 2012
Prostate **prostate adenocarcinoma**	Known - none Suspected -*Cutibacterium acnes* -*Bacteroides massiliensis* -*Streptococcus* spp -*Staphylococcus* spp -STI -Microbial dysbiosis	-Black or African American higher prevalence and mortality rate compared to White. -Altered genetic expression and androgen | estrogen metabolism in African American	Golombos 2017 Cavarretta 2017 Yow 2017	Karakas 2017 Kumar 2018 Sfanos 2013 Wallace 2011
Lung **lung squamous cell and adenocarcinoma**	Know -none Suspected -Microbial dysbiosis -*Chlamydia pneumoniae* -*Granulicatella* spp -*Streptococcus* -HHV-4, HPV, MCV	-Higher incidence and mortality among African American Blacks compared to other racial and ethnic groups -Differential molecular presentation in African American compared to White	Greathouse, 2018	Siegel 2018 Pevsner-Fisher 2016 Heath 2018 Karanth 2018 Mager 2006 Mao 2018 Huffnagle 2017

Table 3 (continued)

Cancer type	Microbial Associations	Racial-related differences	Microbiome Studies	References
Stomach **stomach adenocarcinoma**	Known -*Helicobacter pylori* -HHV-4 Suspected: -Microbiome dysbiosis	-Higher incidence among Asian, Native Hawaiian and other Pacific Islander compared to whites - Hispanics and subpopulations incidence dependent on carrier stage of *Helicobacter pylori*	Cao 2016 Zhang C. 2015 Salyakina 2013	Torre 2016 Siegel 2015
Colorectal **Colon and rectal adenocarcinomas**	Known -Microbial dysbiosis Suspected -E .*Escherichia. coli* -E.*Bacteroides fragilis* -*Fusobacterium nucleatum* -*Campylobacter spp* -*Salmonella* -*Citrobacter* -*Chronobacter* -*Schistosoma japonisium* -HPV	-Higher incidence among African American -African American have higher abundance of Bacteroides and lower abundance of butyrate producing spp.	Marchesi, 2011 Sobhani, 2011 Kostic, 2012, 2013 Castellarin, 2012 Farhana, 2018 Salyakina 2013	Warren 2013 Burnett-Hartman 2008
Liver **liver and intrahepatic bile duct**	Known -hepatitis viruses -parasitic infections Suspected - *Helicobacter pylori*	-Increased incidence among Asian American, Native Hawaiian and Hispanics. Incidence associated with pathogen and behavioral exposures	Cao 2016 Grat 2016	Torre 2016 Siegel 2015 Setiawan 2016

E.= Enterotoxigenic. spp=species

Cancer type	Microbial Associations	Racial-related differences	Microbiome Studies	References
Cervical **cervical squamous cell and endometrial carcinoma**	Know -human papilloma viruses Suspected -*Chlamydia trachomatis* -Microbiome dysbiosis	-Asian, Native Hawaiian, other Pacific Islander and subgroups	Cao 2016	Zhu 2016 Siegel 2015
Head & Neck **Oropharyngeal** **Laryngeal**	Known -HHV-4 -human papilloma viruses Suspected -*Fusobacterium nucleatum* -Microbiome dysbiosis	-Higher incidence among Asian, Native Hawaiian, other Pacific Islander and subgroups -Higher incidence of advanced disease and increased mortality in non-Hispanic black compared to whites	Wang 2017 Cao 2016 Schmidt 2014	Hernandez 2014 Torre, 2016 Gourin 2006 Daraei 2015
Kidney	Known -none Suspected -urinary tract infection associated pathogens	-Native American and Hispanic groups at greater risk with marked geographic differences	Lewis 2013	Batai K, 2018 Li 2014 Bray 2012 Scelo 2017 Pevsner-Fischer 2016

Table lists currently known and suspected microbial agents associated with cancer pathogenesis or have been identified as common causes of infection in cancer patients which disproportionately impact patient outcomes. Asian, Native Hawaiian and other Pacific Islander subpopulations include Asian Indian, Cambodian, Chinese, Filipino, Hmong, Japanese, Korean, Pakistanis, Vietnamese, and those from native origins to Hawaii, Guam, Samoa and other Pacific Islands living in the United States. Hispanic subpopulations include Mexican, Cuban, Puerto Rican, South and Central Americans, Dominican and other ethnic groups from Spanish descent living in the United States. Source: SEER 9 and SEER 13 areas.

American and Asian American compared to their White counterparts (Siegel et al. 2015, Torre et al. 2016). Striking differences are observed within the Hispanic and Asian American ethnic subgroups (Siegel et al. 2015, Torre et al. 2016). Ha et al. (2016) reported that although those of Asian descent have the highest incidence rates of hepatocellular carcinoma, about 36% increase incidence was observed among those of Hispanic origin (Ha et al. 2016). Studies have found that liver cancer disparities are consistent with global variations in hepatitis infection burden and binge drinking among these groups (Siegel, Miller, and Jemal 2018, Global Burden of Disease Cancer et al. 2015, Singh, Siahpush, and Altekruse 2013). Yet, marked differences remain after controlling for known factors which warrants the need to identify potential co-factors in racial disparities (Setiawan et al. 2016). Recently, it has been suggested that gut microbiota may play a role in the development and progression of liver malignancies driven by altered metabolism of nutrients, bacterial metabolites and aberrant DNA methylation (Henao-Mejia et al. 2013, Hattori and Ushijima 2016). In animal studies a distinct altered gut microbiota was associated with progression of liver disease. Xie et al. (2016) found significant alterations in the composition and co-occurrence of several bacterial species including *Aropobium*, *Bacteroides*, *Clostridium* and *Desulfovibrio* species where their composition was correlated with levels of bacterial polysaccharides and pathophysiological characteristics (Xie et al. 2016). In human studies an altered microbiota has been found in liver hepatocellular carcinoma cases with high levels of *Escherichia coli* and other gram-negative bacteria along with reduced levels of *Lactobacillus*, *Bifidobacterium* and *Enterococcus* species (Grat et al. 2016).

2.1.5.1.7 **Cervix**

Cervical cancer is one of the few cancers known to have nearly 100% attributable infectious etiology and marked disparities (Plummer et al. 2016). Incidence and mortality rates have significantly declined between 1992 and 2015 (Siegel et al. 2015, Torre et al. 2016, Siegel, Miller, and Jemal 2018, DeSantis et al. 2016). In the United States there are 13,240 cases (0.8% of all cancer cases) and 4,170 deaths (0.7% of all cancer deaths) each year (SEER data). Still, females living in rural areas have double the risk of dying from cervical cancer than females living in urban areas with definite geographical and racial patterns (Singh 2012, Rauh-Hain et al. 2018, Zahnd, Fogleman, and Jenkins 2018). Females living in rural areas have been found to have 15% to 20% higher mortality rates than their matched counterparts in urban areas (Singh 2012). Persistent racial differences remain among African American, Asian American, Native American and Hispanics. African American and Native American are at greater risk of cancer specific deaths compared to Caucasian females living in rural areas (DeSantis et al. 2016, Zahnd, Fogleman, and Jenkins 2018, Rauh-Hain et al. 2018). Cancer related disparities have been associated to rates of HPV exposure (Zahnd, Fogleman, and Jenkins 2018). Zahnd et al. (2018) assessed rural-urban disparities by cancer stage with racial and ethnic stratification revealing that rural Non-Hispanic

Caucasian females have higher rates of HPV related distant stage cancers. This makes evident the exposure differences that may point to the interplay between biological and environmental influences among racial and ethnic groups and role of co-factors. Cofactors may enhance susceptibility of certain populations to cervical cancer after HPV infection. *Chlamydia trachomatis* co-occurrence in cervical cancer for example, has been found to be an independent predictor and carcinogenic co-factor for racial disparities for the observed differences in urban vs. rural dwellers (Zhu et al. 2016).

2.1.5.1.8 **Head & Neck**

Squamous cell carcinomas of the head and neck are a heterogeneous group of cancers categorized by the area in which they begin collectively known as head and neck squamous cell carcinoma (cancer.gov). It is the ninth most common cancer in the United States (Siegel, Miller, and Jemal 2018). It accounts for 3% of new cancer cases with marked racial differences in mortality and late stage presentation (Siegel, Miller, and Jemal 2018). African American have the highest distant rate for HPV associated oral cancers (Zahnd, Fogleman, and Jenkins 2018). Several factors contribute to observed racial differences including socioeconomic status, access to care, exposures and biological factors (Daraei and Moore 2015, Ragin et al. 2016). Differences in outcomes may be attributed to a combination of these (Ragin et al. 2016). Recent studies have identified HPV as an etiological agent for oropharyngeal and invasive laryngeal cancers (Tanaka et al. 2018, Hernandez et al. 2014). Infection status has been suggested as a biomarker in the interplay of racial disparities. In a study of 87 African American head and neck cancer patients and 261 matched age and smoking status white patients, Ragin et al, found that when stratifying by HPV status significant differences remained among African American and their white counterparts for tumors of the larynx (HR 3.36 95%CI 1.62-7.0) after adjusting for socioeconomic status and other confounding variables (Ragin et al. 2016). Contributions from other microbes have been suggested. Changes in microbial oral communities have been postulated as risk factors and can be used as biomarkers for pre-cancerous lesions and cancer progression (Schmidt et al. 2014, Wang et al. 2017) Microbial shifts have been detected where significantly reduced Firmicutes and Actinobacteria phyla have been observed in tumor tissue compared to adjacent normal or non-cancerous normal controls by several teams. Similar microbial associations have been observed at the phylum level in Chinese and American population samples and it remains to be seen if inter-individual racial variation lays at the genus or species level (Bik et al. 2010).

2.1.5.1.9 **Kidney**

Native American and Alaskan Natives have higher incidence and mortality rates of renal cancers than other racial or ethnic groups (Batai et al. 2018, Li et al. 2014). Despite overall reduction in renal cancers among Caucasian, rates for Native American and Alaskan Natives have remained steady (Li et al. 2014). Li et al. (2014) reported that between 2001 and 2009 incidence rate significantly increased for Native American and Alaskan Natives compared to Caucasian with an annual percent change (APC) of 3.5% versus 2.1% per year respectively (NA/AN 95%CI =1.2, 5.8; white 95%CI = 1.4,2.8) (Bray et al. 2012).

Hispanics also suffer from a disproportionate burden of renal cancers with striking geographic-regional differences. The burden is also higher among men for both Native American and Alaskan Natives and Hispanics populations (Batai et al. 2018). Risk differences between ethnic groups may be explained by common attributable factors including behavior, socioeconomic status, obesity, smoking and presence of comorbidities (diabetes or cardiovascular disease), factors that have been linked to microbial dysbiosis. However, sex and treatment response differences cannot be explained by the same factors and remain a mystery with few hypothesis on the potential microbial effects on immunosurveillance mechanisms and therapeutic response (Scelo et al. 2017, Pevsner-Fischer et al. 2016). In a retrospective study, prior urinary tract infection was associated risk of developing malignancy with similar proportions on renal cancer incidence related to behavioral exposures (smoking and alcohol consumption) and sex differences in those infected versus non-infected history (Parker et al. 2004). Although there is limited evidence on the microbial compositional structure of the renal cancer tumors and the potential long term effects, several studies have now identified differences in microbiota of the urinary tract of healthy individuals that may impact renal cancer pathogenesis and help explain differences in sex and race (Lewis et al. 2013).

2.1.6 **Summary**

Cancer is the second leading causes of death in the United States. One in five cancer cases is attributed to infectious agents around the world. The divergent relationship of tumor microbiota in cancer pathogenesis was discussed along the potential impact of microbiological agents on racial related disparities in cancers of the breast, prostate, lung, colorectum, liver, cervical, head and neck, and kidney. A considerable proportion of these cancer cases are of unknown etiological that underscore the importance of understanding the interplay of contributory factors such as microbial contributions. Recent studies have evaluated the microbial content and its influence in cancer pathogenesis. Few have identified racial related differences attributable to microbial dysbiosis. Given the close interaction of microbes with host immune reaction and the metabolite degradation “it is naïve to believe that our microbiome has no impact carcinogenesis” and thus possibly in the observed racial variations which can benefit or negatively impact people from certain racial and ethnic groups. Examination of the potential role of microbes is pivotal to developing new prevention and treatment strategies to reduce and eliminate racial disparities

2.2 **Identification of Racial-Related Microbial Differences across Cancers Derived from Human Sequencing Data**

2.2.1 **Abstract**

Microbiological infections account for up to 25% of the total global cancer burden, one of the leading causes of morbidity and mortality worldwide. Racial-disparities persist in infection-associated cancers. Underlying causes for racial disparities are multifactorial and not all well understood. Recent evidence points out the potential role of microbial composition in racial-related disparities. While much effort has gone into the characterization of the gut microbiota, the microbial compositional differences of tumor tissue is less explored. Identification of tissue-associated microbial differences is challenging and computationally intensive. New opportunities are becoming available with newer bioinformatics tools and merged data networks. Computational frameworks can assist in the interpretation of the microbial impact on tumor tissue, cancer pathogenesis and possible roles behind racial-related cancer disparities. Here we review current computational frameworks designed to derive microbial information from host-sequencing data, along with a brief discussion of post-processing tools which can inform our understanding of racial-related differences in cancer.

2.2.2 **Introduction**

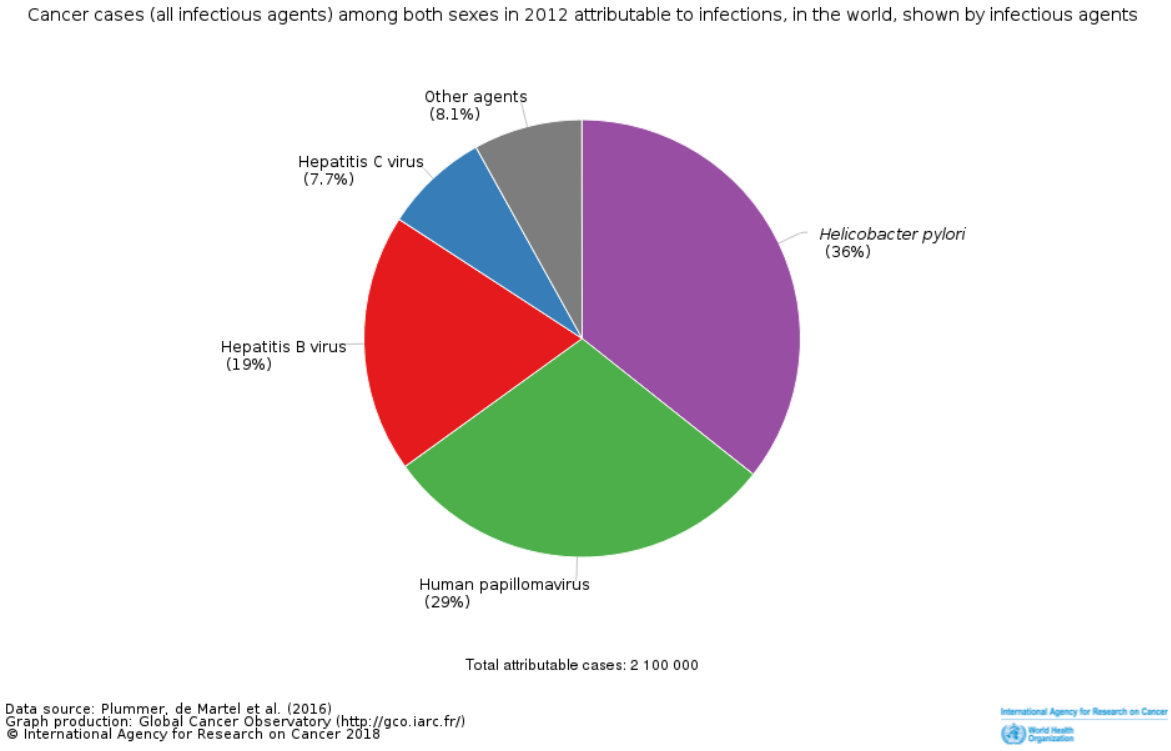
Cancer is one of the leading causes of morbidity and mortality worldwide. Annually an estimated 18.1 million are diagnosed and 9.6 million die from the disease globally (WHO 2018). There is evidence that infection may contribute to cancer burden and cancer disparities (Siegel et al. 2015, Torre et al. 2016, Deshmukh et al. 2017, Wallace, Martin, and Ambs 2011). Determinants of racial-related disparities are complex involving environmental, social, behavioral, cultural and biological cofactors impacting development, progression and clinical outcomes of cancers. Recent evidence highlights the importance and possible role of the altered microbiota in cancer development, treatment response, and racial disparities (Hattori and Ushijima 2016, Goyal et al. 2016, Pevsner-Fischer et al. 2016). Unbiased evaluation of microbial patterns across various cancer types may elucidate their role. Characterization of the microbiota and its impact on racial-related disparities is challenging and computationally intensive. These can be simplified by integrated analyses approach incorporating epidemiological and clinical data with microbial compositional characterization, as well as post-process identification of functional relevant alterations. Here we briefly review bioinformatics frameworks designed to derive microbial information from host-sequencing data and post-processing pipelines designed to identify functional prediction of the tumor microbiota as part of integrated approach to understand microbiota impact on racial related differences.

2.2.3 **Infectious disease burden and cancer disparities**

According to the World Health Organization (WHO), microbiological infections account for approximately 25% of the total global cancer burden (WHO 2018). Infection-associated cancers are commonly attributed to viruses including human herpes virus 4, hepatitis B virus, hepatitis C virus, human immunodeficiency virus 1, human papilloma virus, human T-lymphotropic virus and herpes virus (IARC 2012). Parasitic infections casually linked to cancers include *Clonorchis sinensis*, *Opisthorchis viverini* and *Schistoma haematobium* liver flukes (IARC 2012). Only one bacterial agent, *Helicobacter pylori,* is affirmatively linked to cancer although the burden is substantial with over 1/3 of infection related cancers attributed to the bacterium (de Martel et al. 2012, Bouvard et al. 2009, Cogliano et al. 2011). **Figure 5**. A perturbation to the microbial community or the microbiome-host relationship, known as dysbiosis, is widely recognized to be a contributing factor in the promotion of disease including cancer (Human Microbiome Project 2012, Schwabe and Jobin 2013). Microbial contributions to racial-related disparities in cancer pathogenesis are not well understood. Evidence points to possible contribution of infection burden to cancer disparities (Sobhani et al. 2011, Zhu et al. 2016, Hernandez et al. 2014, Setiawan et al. 2016). Therefore, characterization of microbial profiles of precancerous and cancerous tissue may revolutionize our understanding of cancer pathogenesis and disparities.

Understanding the complexity of the interactions between commonly identified racial disparities contributing factors and exposure to cancer-modulating microbial organisms are key to reducing and eliminating racial related disparities. Several studies have evaluated the microbial content and its influence in cancer pathogenesis, most have concentrated in gut and oral microbiome, and few have identified racial related differences attributable to microbial dysbiosis (Contreras et al. 2016, Schmidt et al. 2014, Farhana et al. 2018). To our knowledge no work to date, addressing racial differences in the microbial composition within the tumor and adjacent tissue microenvironment has been conducted. In fact, in recent high throughput sequencing microbial analyses, aside from the histopathological descriptions, most do not describe patient population demographics beyond age and sex if mentioned**.** **Table 4**. Two teams, Cao et al. (sequencing data) and Farhana et al. (colonic effluent) have described racial differences in their work. Given the close interaction between microbes and the host responses it is important to clearly identify compositional structure and clinically relevant functional pathways with an integrated approach. Interpretation of compositional difference contributions is facilitated by new sequencing technologies and various computational frameworks some which we review briefly here.

Figure 5 *Proportion of infectious agents responsible for cancer incidence in both cases worldwide*



Cancer cases (all infectious agents) among both sexes attributable to infections shown by infectious agents' proportion of attributable fraction. *Helicobacter pylori*, a bacterium, is attributed to1/3 of the total global burden. Data source: Plummer, de Martel et al. (2016). Chart created through GLOBOCAN interactive (gco.iarc.fr).

2.3.4 **Microbial detection in high throughput sequencing data**

Next Generation Sequencing (NGS) also known as high-throughput sequencing technologies provide a powerful tool for the evaluation of the role of microbes in cancer development and progression as well as disparities across populations. NGS is a useful and unbiased tool for the identification of previously undetected or unsuspected causative microorganisms in molecular diagnostics (Daly et al. 2015). It has become vital and necessary to the integrative analysis of cancer biology enabling description of the mutational and molecular landscape of cancer (Reuter, Spacek, and Snyder 2015). This technology, however, produces relatively short reads that limits interpretation. NGS has driven the development of new alignment algorithms designed to handle short reads, which in turn has revolutionized the understanding of the relationship between genomic variation and phenotype (Reuter, Spacek, and Snyder 2015, van Dijk et al. 2014). Integrative techniques take advantage of the production of short reads and predominance of host-derived sequences to examine pathogen-host interaction including their correlation with metabolic and regulatory mechanisms in cancer (Marchesi et al. 2011, Warren et al. 2013, Schmidt et al. 2014, Contreras et al. 2016, Arthur et al. 2014, Cristescu et al. 2015, Wang et al. 2017). Several studies have examined microbial comminutes (tumor microbiome, gut microbiome, oral microbiome,

virome and bacteriome) in cancer populations by high throughput sequencing methods. **Table 4.** Taken together these studies have provided much needed information about microbial diversity, richness and abundance variations, across healthy and cancer states. In addition, they afford the opportunity to evaluate microbial profile and its correlation to the clinical features. Although establishment of a causal relationship requires a more detailed characterization of the human microbiome and microbial population dynamics; host-microbial sequencing and integration of clinical and epidemiological data can provide valuable information to the understanding of the role microbiota plays in cancer pathogenesis and cancer disparities.

2.2.5 **Racial disparities in high throughput sequencing data**

Racial and ethnic groups are underrepresented in some sequencing efforts. Underrepresentation of ethnic minorities may also impact interpretation of results from sequencing data. The low representation of racial minorities in sequencing efforts may limit the ability to discern population differences in mutational frequency (Spratt et al. 2016). According to Spratt et al. (2016) the ability to detect mutations in a subpopulation with enough power, depends on the mutational frequency of the background and the mutational rate of the target as well as the absolute sample size (Spratt et al. 2016). Recently, Zhang et al. (2017) estimated racial disparities in cancer incidence rate, the effect of race on survival and mutation burden within TCGA cancer cohorts (Zhang et al. 2017). Racial disparities in cancer incidence, survival, and/or tumor mutation burden were observed for bladder urothelial carcinoma, glioblastoma multiforme, head and neck squamous cell carcinoma, kidney renal clear cell and papillary carcinoma, lung adenocarcinoma and squamous cell carcinoma, ovarian serous cystadenocarcinoma, uterine corpus endometrial carcinoma, colon adenocarcinoma, hepatocellular carcinoma and stomach adenocarcinoma (Zhang et al. 2017). Similarly, ability to detect microbial differences across racial groups depends on the infection and prevalence patterns of the microbe in the background population. This highlights the importance of genomic and clinical epidemiological data integration across institutions, particularly of catchment areas. Validation against enriched population-based data from catchment areas and clinical trials are equally important for meaningful interpretation of microbial detection in cancer particularly when looking to address racial-related disparities. Information on race and ethnicity from integrated data networks, are typically obtained from self-reported measures, which may introduce bias. In this case, the use of ancestral biomarkers may be appropriate to validate self-reported racial/ethnic information and understand impact of differential patterns between background and target population on racial disparities. Data integration across platforms and multiple study efforts minimize bias and strengthens results.

Table 4 *Summary of microbial detection in high throughput sequencing data*

Author	Journal	Year	Cancer	Microbial agent	Data Type	Sample Type	Paired Method? Included (N)	Additional information	Racial Diff.?
Marchesi et al.	PLOS One	2011	Colorectal	Bacteriome	16S	Tissue	▪ Paired 6 tumor : 6 adjacent	▪ Examined dysbiosis	No
Shobhani et al.	PLOS One	2011	Colorectal	Gut microbiome	16S	Stool	▪ Non-paired 60 case: 119 normal	--	No
Kostic et al.	Genome Res	2012	Colorectal	Microbiome	WGS	Tissue	▪ Paired 9 tumors: 9 adjacent	▪ Validated against TCGA and animal model	No
Castellarin et al.	Genome Res	2012	Colorectal	Bacteriome	RNA-seq	Tissue	▪ Matched 11 tumor: 11 normal	--	No
Riley et al	PLOS Comp Biol.	2013	TCGA	Bacteriome	RNA-seq	Seq. data	▪ Paired 632 tumor: 220 normal	▪ Examined lateral gene transfer ▪ TCGA and 1000 GP ▪ 10 TCGA cohorts	No
Tang et al.	Nature Comm.	2013	TCGA	Virome	RNA-seq	Seq. data	▪ Paired 4433 tumor: 404 normal	▪ Viral integration. ▪ 19 TCGA cohorts	No
Warren et al.	Microbiome	2013	Colorectal	Bacteriome	RNA-seq	Tissue	▪ Paired 65 tumor: 65 matched normal	▪ Examined co-occurrence & host gene expression	No
Khoury et al.	JVI	2013	TCGA	Virome	RNA-seq	Seq. data	▪ Non-paired 3775 tumors	▪ 18 TCGA cohorts included	No
Kostic et al.	Cell	2013	Colorectal	Microbiome	RNA-seq	Seq. data	▪ Non-paired 133 tumors	▪ Also performed qPCR in tissue & stool ▪ Animal work	No
Salyakina et al.	Human Genomics	2013	TCGA	Virome	RNA-seq	Seq. data	▪ Paired 958 RNA tumors 120 DNA tumor 91 DNA normal blood 22 DNA normal tissue	▪ 9 TCGA cohorts included ▪ Computational pipeline development	No

Table 4 Summary of microbial detection in high throughput sequencing data (continued)

Author	Journal	Year	Cancer	Microbial agent	Data Type	Sample Type	Paired Method? Included (N)	Additional information	Racial Diff.?
Schmidt et al	PLOS One	2014	Oral	Bacteriome	16S	Swab	▪ Paired 5 lesion: 5 anatomically matched	▪ Validation with swabs from cancer, pre-cancer lesions & healthy controls	No
Xuan et al.	PLOS One	2014	Breast	Bacteriome	16S	Tissue	▪ Paired+ 39 tumor: 39 adjacent 29 controls	▪ Examined expression profiles	No
Zhang et al.	Genome Biology	2015	Gastric	Microbiome	WGS	Tissue	▪ Non-paired 29 gastric cases	▪ Computational pipeline development validated against HMP and TCGA	No
Banerjee et al.	Scientific Reports	2015	Breast	Microbiome	Probe Array	Tissue	▪ Matched 100 cases: 17 matched: 20 non-matched	--	No
Chan et al.	Scientific Reports	2016	Breast	Bacteriome	16S	Aspirate	▪ Non-paired 25 history: 23 healthy controls	▪ Skin swabs also evaluated. ▪ Functional pathway prediction completed	No
Thomas et al.	Frontiers CIM	2016	Rectal	Bacteriome	16S	Tissue	▪ Case-control 18 case 18 control	--	No
Iyer et al	Infectious Agents & Cancer	2016	Gallbladder	Microbiome*	WXS	Tissue	▪ Matched 17 tumor: 9 normal	▪ Examined co-infection with HPV and P53 status	No
Cao et al.	Scientific Reports	2016	TCGA	Virome	RNA-seq	Seq. data	▪ Paired 6813 tumor: 559 adjacent	▪ Computational pipeline development. ▪ Examined integration & strain variants	Yes
Yu et al.	Int. J Cancer	2017	Gastric	Bacteriome	16S	Tissue	▪ Paired 80 tumor: 77 non-malignant	▪ Functional modules	No
Robinson et al.	Microbiome	2017	TCGA	Bacteriome	RNA-seq	Seq. data	▪ Non-paired 1581 tumor: 284 normal	▪ Examined contamination ▪ Included blood normal	No
Thompson et al.	PLOS One	2017	Breast	Bacteriome	RNA-seq	Seq. data	▪ Paired 668 tumor: 76 adjacent	▪ Gene expression profile	No

Table 4 Summary of microbial detection in high throughput sequencing data (continued)

Author	Journal	Year	Cancer	Microbial agent	Data Type	Sample Type	Paired Method? Included (N)	Additional information	Racial Diff.?
Cavarretta	EAU	2017	Prostate	Bacteriome	16S	Tissue	▪ Paired 16 tumor: 16 peritumor 16 non-tumor	--	No
Zhao et al.	Scientific Reports	2017	Oral	Bacteriome	16S	Tissue	▪ Paired 40 lesion: 40 anatomically matched	▪ Functional modules	No
Golombos et al.	Oncology	2017	Prostate	Gut microbiome	16S	Stool	▪ Case-control 12 case: 8 control	▪ Functional modules	No
Yow et al.	Infectious Agents & Cancer	2017	Prostate	Microbiome	16S	Tissue	▪ Non-paired 20 aggressive tumor	▪ Total RNA sequencing comparison	No
Wang et al.	Genome Medicine	2017	Oral	Oral microbiome	16S	Tissue	▪ Paired 121 tumor: 121 adjacent	--	No Race described
Cantalupo et al.	Virology	2018	TCGA	Virome	RNA-seq	Seq. data	▪ Paired 3074 tumor: 1488 normal	▪ Computational pipeline development. ▪ Utilized WGS and WXS files in addition to RNA-seq. ▪ 22 TCGA cohorts	No
Farhana et al.	WJGP	2018	Colorectal	Bacteriome	16S	Effluent	▪ Non-paired 52 AA: 46 CA CRC patients	▪ Examined microbial dysbiosis	Yes
Gopalakrishnan et al.	Science	2018	Melanoma	Gut microbiome	16S	Stool	▪ Non-paired 30 responders: 13 non-responders	▪ Examined immune profiling. ▪ Buccal, tumor and blood specimens also collected at different time points ▪ WGS performed in subset	No

Table 4 Summary of microbial detection in high throughput sequencing data (continued)

Author	Journal	Year	Cancer	Microbial agent	Data Type	Sample Type	Paired Method? Included (N)	Additional information	Racial Diff.?
Hayes et al.	JAMA Oncol	2018	Head & Neck	Oral microbiome	16S	Oral wash	Nested case-control 129 cases: 254 controls	▪ Examined cancer risk	No
Greathouse et al.	Genome Biology	2018	Lung	Bacteriome	16S	Tissue	Non-paired 144 cases 33 controls	▪ Validation against TCGA tumor and adjacent (974/108) ▪ Discussed P53 mutation associations	No; Race described

Table shows studies evaluating microbial comminutes in cancer populations by high throughput sequencing methods. Microbiome=refers to studies evaluating bacteria, virus and/or others [gut: mostly from stools, oral: mostly swabs or saliva products]; Bacteriome= refers to bacterial community; Virome= viral community

2.2.6 **Computational frameworks and microbial detection in cancer**

Data mining high throughput sequencing data using bioinformatics tools and methods provide great opportunities in understanding the role of the microbiota in cancer racial-related differences. Bioinformatics computational frameworks are able to accommodate user defined parameters and deliverables to better understand the basis of biological concepts (Leipzig 2017). Numerous state-of-the-art bioinformatics tools and methods from data collection to analyses are available today that support identification of microbial novel targets in cancer diagnostics, treatment, prevention and control. Several studies have demonstrated that pathogenic and commensal microbes can be derived from human cancer tissue sequences utilizing various bioinformatics computational approaches and sequential filtering steps (Daly et al. 2015, Isakov, Modai, and Shomron 2011, Weber et al. 2002, Xu et al. 2003, Kostic et al. 2011, Tae et al. 2014, Salyakina and Tsinoremas 2013). Pathogen detection derived from human sequences has been primarily completed by one of three approaches*, reference-based* or *reference-free*, or *mixed-methods* with one fundamental core pipeline involving the removal of human-host to characterize remaining sequencing reads. Pathogen detection algorithms can be classified by the methodology, meaning the order in which the human sequencing reads are identified or removed, whether the human reads are removed before or after extracting pathogen reads, and what happens to remaining unmapped reads. **Figure 6**. Here, we review bioinformatics computational frameworks designed to identify microbiota derived from human sequences by computational subtraction, marker-based filtration, or mix-methods approaches with applications in human cancer. Computational frameworks that strictly match sequencing reads to pathogen libraries or those designed for direct metagenomics analyses are not included. See Noecker et al. and Nooij et al. for an in depth reviews of these tools (Nooij et al. 2018, Noecker et al. 2017).

Computational subtraction methods for microbial identification and discovery derived from human tissue sequences were first introduced about 15 years ago (Xu et al. 2003, Weber et al. 2002). These early approaches involved creation of cDNA library and sequential subtraction of human-expressed sequence tags to derive pathogenic and commensal organism information from human disease and were computationally intensive. Newer methods take advantage of high throughput data repositories' unmapped-to-human sequences and have lower computational requirements. In high throughput sequencing, about 10% of the reads are flagged unmapped to the human genome after alignment, which could in part belong to the human microbiome or yet uncharacterized human genome (Tae et al. 2014). In fact, this characteristic shortfall provides the basis to multiple bioinformatics approaches (**Table 5**) which help better understand microbiota differences in infection-associated cancers.

2.2.6.1 Reference-Based: Computational subtraction methods are mostly reference-based approaches. Reference-based by definition require mapping to a reference, in this case human host genome, then allocating all leftover unmapped-to-human reads to pathogen target genomes. Bioinformatics computational frameworks such as PathSeq (Kostic et al. 2011) and SRSA, short RNA subtraction and assembly (Isakov, Modai, and Shomron 2011), consider unmapped-to-human sequencing reads as input data and are able to lower computational costs while facilitating novel discoveries.

Figure 6 *Generic pipeline of computational frameworks designed to identify microbial reads from human derived sequences*

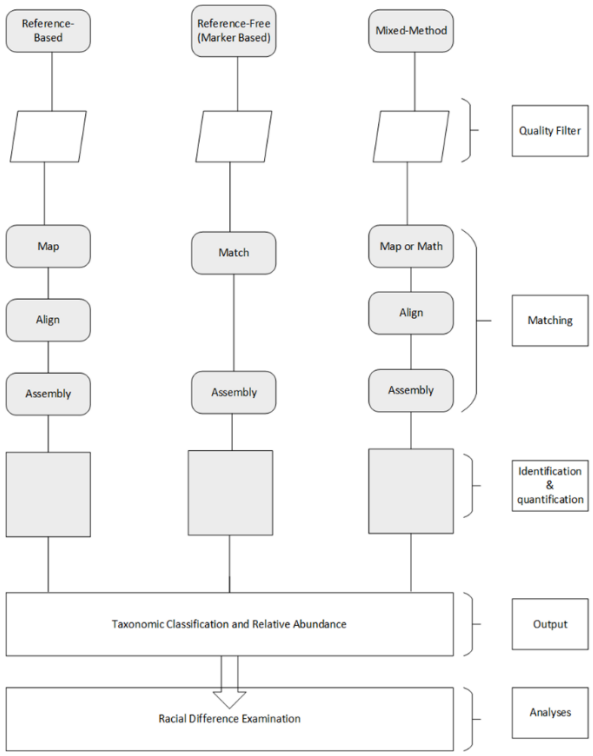


Figure shows generic pipeline divided into 5 general steps, quality filter pre-process, map, alignment and assembly match (match processing), identification & classification (quantitation processing), microbial output processing, and microbial analyses with integrated epidemiological and clinical data . During quality filter, sequencing reads are trimmed, during the processing steps, reads are mapped and aligned to either human or pathogen reference sequences or other key identifying factor before the final steps of identification and analyses of identified sequences. The final steps may involve correlation and functional relevant identification of molecular pathways based on the identified sequences.

PathSeq combines alignment and de novo assembly with a two-pass subtraction process (Kostic et al. 2011). It aligns the sequencing reads to target genomes and quantify their abundance based on the total number of aligned sequencing reads and the genome coverage, enabling identification of both commensals and pathogens whether known or novel (Kostic et al. 2011). However, the two-pass filtration process may eliminate a high number of sequences, which can increase filtration costs and limit identification. PathSeq has been utilized in pathogen identification for various infection-associated and

inflammation associated cancers types, notably yielding the association between *Fusobacterium nucleatum* with colorectal cancer and *Bradyrhizobium enterica* in cord colitis (Kostic et al. 2012, Bhatt et al. 2013). SRSA utilizes short RNA mapping and assembly to identify pathogens in relation to host sequencing reads (Isakov, Modai, and Shomron 2011). SRSA has capability for use in microbial identification in infection-associated cancers. However, initial work was limited to mycoplasma detection in HIV-1 cell lines and its computational methods are also not freely available. Unlike SRSA, CaPSID, computational pathogen sequence identification (Borozan et al. 2012), is web-based and an open source platform, that similar to PathSeq, performs mapping and de novo assembly. CaPSID differs in its single-pass alignment and filtration process where both human and pathogen reads are aligned to reference genomes while separating those that do not match either for de novo assembly simultaneously (Borozan et al. 2012). Its potential in cancer microbial differential analyses was demonstrated by Borozan et al. (2012) in STAD-TCGA and stomach adenocarcinoma samples from other cancer networks (Borozan et al. 2018). Borozan et al. (2018) evaluated HHV-4 variants to determine oncogenic potential differences among samples from different country origins providing evidence of the potential for future studies in HHV-4 ethnic and racial differences in gastric cancer.

Similar reference-based computational subtraction methods with potential in cancer microbiome and racial differences downstream analyses include PathoScope 2.0 (Hong et al. 2014), SURPI, sequence-based ultra-rapid pathogen identification (Naccache et al. 2014), VirusScan (Cao et al. 2016) and MetaShot (Fosso et al. 2017). Unlike PathSeq, SRSA and CaPSID, PathoScope 2.0 does not perform de novo assembly, instead it utilizes penalized statistical mix-model and probabilistic pathogen identification (Hong et al. 2014). It also provides detailed reports with core and optional module format that enable user customization. A target reference genome must be present for precise identification of microbes. PathoScope is designed to identify low abundant strains, making it an ideal tool for host derived microbial analyses. PathoScope is able manage the low abundance relative to human host microbial reads in sequencing data. Zhang et al. (2015) incorporated PathoScope 2.0 methods for relative abundance estimation with its WGS PathSeq-based microbial detection pipeline in gastric cancer clinical samples, Human Genome Project derived and TCGA sequencing samples (Zhang et al. 2015). SURPI was also designed for pathogen detection from clinical samples for surveillance similar to PathoScope 2.0 with the capacity for quantitative and semi-quantitative identification; meaning it can perform mapping and de novo assembly for divergent microbial analyses (Naccache et al. 2014). SURPI has been validated against samples from colon and prostate cancer derived datasets. VirusScan is also a referenced-based computational subtraction approach designed to profile viral composition, abundance and integration sites in human tumors utilizing unmapped-to-humans and poorly mapped to human genome reads (Cao et al. 2016). Cao et al. (2016) identified racial-related differences of viral strains within two TCGA cancer

Table 5 *Computational frameworks designed to detect microbiota from human sequences by subtractive, filtration or mixed methods*

Framework	Dependency	Approach	3rd Party Tools	Input | Output	Advantages/Disadvantages	Cancer Validation	References
PathSeq	Reference-based	-Alignment & *de novo* assembly	BLAST BLASTN BLASTX MAQ MegaBLAST RepeatMasker Velvet	Input: -RNA-seq or DNA-seq Output: -Pathogen presence/absence	-Scalable cloud computing -Feasible for known and novel pathogen identification -Two-pass subtraction with increased filtering costs	-Cervical cancer (cell line and simulated data) -TCGA ovarian	Kostic, 2011 Bhatt, 2013
SRSA	Reference-based	- Alignment & *de novo* assembly	Velvet MegaBLAST BLAST BWA TopHat	Input: -RNA-seq Output: -Species level taxonomy characterization (prevalence)	-Incorporates sample pre-processing, quality filtering, sequence mapping and assembly -Not freely available -No known updates -Original work validation was limited to cell line.	HIV-1 cell line	Isakov, 2011
CaPSID	Reference-based	-Mix-method, simultaneous alignment, filtration & *de novo* assembly	BioPython Bowtie2 Trinity	Input: -RNA-seq or DNA-seq Output: -Top-hit pathogen genome identification ranked by maximum gene coverage	-Web-based, open source and scalable application -Modular analyses -Single pass filtering, which may fail to subtract host reads	-Ovarian cancer -TCGA stomach	Borozan, 2012, 2017
SURPI	Reference-based	-Dual scanning mode -Known pathogens identification or *de novo* assembly	SNAP RAPSearch BWA BLASTN Bowtie2 DUST in PRINSEQ	Input: -Paired-end metagenomic Output: -Species level taxonomic classification and coverage map	-Scalable to cloud or standalone servers -Capacity to incorporate reference database -Dual mode: quantitative and semi quantitative pathogen identification	-Prostate cancer (cell line, tissue biopsies) -Colorectal cancer (tissue biopsies)	Naccache, 2014

Framework	Dependency	Approach	3rd Party Tools	Input | Output	Advantages/Disadvantages	Cancer Validation	References
PathoScope 2.0	Reference-based	-Penalized probabilistic identification -Modular filtration, alignment and assignment	SAMtools BLASTX Bowtie2 thetaPrior	Input: -Metagenomic or genomic (RNA-seq or DNA-seq) Output: Strain level pathogen relative abundance	-Modular detailed result reporting with -Designed for low abundance strain level identification -MySQL server required; no connection to population structure of relevant species	-TCGA stomach	Hong, 2014 Zhang, 2015
VirusScan	Reference-based	-Identification of known viral and integration sites	BWA BLAST MegaBLAST Pindel RepeatMasker PHYLIP	Input: RNA-seq Output: Viral read abundance and integration sites	-Designed for viral identification -Abundance and integration sites analyses	-TCGA cancer cohorts	Cao, 2016
MetaShot	Reference-based	-Two-step similarity filtering and taxonomic assessment	Bowtie2 TANGO STAR Bash	Input: RNA-Seq or DNA-Seq Output: -Assigned read report and Krona plot with relative abundance	-Extracts unassigned reads -Allow for functional annotations -Slower than other applications	None	Fosso, 2017
ConStrains	Reference-free	-Marker-based (SNP patterns) -Strain-level prediction	MetaPhlAn, PhyloPhlAn, Bowtie2, SAMtools, Metropolis-Hasting Monte-Carlo	Input: -Metagenomics (RNA-seq) Output: -Strain-level prediction and relative abundance	-Single reference strain collection -Facilitates functional analyses when combined with reference genome-based gene coverage metadata	None	Luo, 2015

Framework	Dependency	Approach	3rd Party Tools	Input | Output	Advantages/Disadvantages	Cancer Validation	References
RINS	Reference-based	-Intersection based identification and removal	Bowtie BLAST BLAT Trinity	Input: -Mate-paired RNA-seq unmapped reads Output: Pathogen contigs	-Requires prior knowledge of reference -Detection limited to user defined parameters	-Prostate cancer (cell line)	Bhaduri, 2012
GRAMMy	Reference-based	-Mix- model Bayesian, Expectation Maximization & maximum likelihood estimation	BLAST BLAT MAQ Bowtie PerM BLASY	Input: -Metagenomics reads Output: Genomic relative abundance as numerical vectors	-User flexibility -Probabilistic handling of ambiguous hits -Computational efficient.	None	Xia, 2011

Table displays computational workflows designed to derive microbial content from human sequences by subtractive and filtering methods broadly categorized as reference-based, reference-free and mix-method which have been used in recent cancer microbial analyses with potential use in racial differences studies

cohorts, LIHC and STAD, utilizing VirusScan. MetaShot is similar to prior mentioned reference-based approaches in that it shares a two-step filtration method to identify candidate pathogens; however, is a bit more stringent in its taxonomic assignment (Fosso et al. 2017). This feature enables functional annotation with great potential in racial disparities studies. On the other hand, its stringent approach comes with higher computational costs and has yet to be validated in cancer datasets.

2.2.6.2 Reference-Free: Other methods like marker-based approaches utilize pre-defined target genomic markers like k-mers, single nucleotide polymorphisms (SNP), or unique tag libraries to identify and retain pathogen information while removing human host sequences from further consideration. Reference-free, marker-based approaches such as ConStrains, conspecific strains (Luo et al. 2015) rely on the creation of SNP profiles to predict pathogen strains contained within the sequencing sample. However, approaches such as this are not completely reference-free, rather minimally reference-dependent (Luo et al. 2015). ConStrains infers microbial abundance of conspecific strains utilizing a SNP patterns and de novo assembly with prediction estimation based on Metropolis-Hasting Markov Chain Monte-Carlo model. Although ConStrains has not been used in cancer genomic data, it has the capability for functional analyses which are pivotal in understanding divergent microbial effects in cancer particularly those of infectious etiology.

2.2.6.3 Mixed-method approaches: Mix-methods can be reference-free, like intersection analysis or reference-based like in mixture-model approaches, which utilize both reference and marker-based methods. Mixture model approaches differ from traditional computational subtraction in that these either map against a pre-determined pathogen reference in sequence (Bhaduri et al. 2012, Fosso et al. 2017), against both human and pathogen in parallel (Naeem, Rashid, and Pain 2013), or some combination (Xia et al. 2011) of these before filtering out human host sequences. RINS, rapid identification of non-human sequences (Bhaduri et al. 2012), uses intersection analysis and pre-defined query reference that include genomes of viruses, bacteria or any other pathogen rather than first mapping to the human reference genome. RINS was validated in prostate cancer and has low computational requirements, but can only detect the pathogens that are explicitly defined within the query reference. This risks removal of unknown sequences and hinders novel pathogen discovery. Mixture model approaches utilize expectation maximization algorithms to calculate genome relative abundance of non-host microbial sequences to obtain meaningful data of the relative abundances. GRAMMy, genome relative abundance estimation framework using mixture model theory (Xia et al. 2011), is a mixture model approach designed to use either mapping or de novo assembly in the absence of a reference genome for relative abundance estimation at different taxonomic levels. These approaches facilitate identification of pathogens within the tumor microenvironment directly from human sequences and streamline functional prediction and correlation with epidemiological and clinical features of the population.

2.2.7 **Computational pipelines and functional prediction of microbial differences**

Recent work in gut microbiome have revealed the utility of taxonomic differences, epigenetic, heritable and co-occurrence patterns in the understanding of racial disparities (Brooks et al. 2018). Microbial compositional differences and racial variations have been thoroughly reviewed in Gupta et al. (Gupta, Paul, and Dutta 2017). From these and other works we understand that accurate interpretation of microbial impact cancer disparities involve more than compositional differences across racial and ethnic groups. Functional annotation and prediction of molecular process are equally important in the identification of clinically relevant microbial interactions in the human host. Many post-processing tools have been developed to translate microbial compositional outputs of the before mentioned tools into predicted mechanisms through which microbiome may influence host immune responses, gene expression and protein expression. For example pipelines such as PICRUSt (Langille et al. 2013), Tax4Fun (Asshauer et al. 2015), and ShortBRED (Kaminski et al. 2015) can assist in the identification of functional annotations and in the description of subtle differences across racial and ethnic groups. Although these tools are designed to predict functional profiles from 16S rRNA gene derived metagenomic data, they have application in human genome sequencing derived microbial profiles when used in integrated approaches. PICRUSt (phylogenetic investigation of communities by reconstruction of unobserved states) pipeline infers microbial community host-associated functional composition based on gene annotation databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) reference genomes or the Clusters of Orthologs Group (COGs) (Kanehisa et al. 2002). Similarly, Tax4Fun predicts the functional capabilities of microbial communities based on 16S rRNA datasets which provide a good approximation to functional profiles obtained from metagenomic shotgun sequencing approaches and has been successfully used to identify signs of ethnic acculturation in oral microbiota (Hoffman et al. 2018). ShortBRED quantifies abundance of functional gene families and predicts protein profiles within the sample including antibiotic resistance and virulence factors protein families which are pivotal in the understanding of outcome racial-related differences.

2.2.8 **Summary**

There is great diversity present in the human tumor microenvironment that makes identification of the microbial community challenging. Sequencing technologies and use of these computational tools permit the discovery of new microbes that are non-culturable and would otherwise remain undiscovered (Relman 1998). Profiling and characterization of the microbial community and functional annotations can provide information on the effects of microbiota on colonized tissue, progression of inflammation, alteration of cellular processes and effects on tumor promoting genes within the tumor microenvironment. Computational frameworks for microbial detection evaluated here are broadly classified as reference-based or reference-free and mainly utilize computational subtraction, marker-based or mixture approaches. Here, we reviewed few bioinformatics computational frameworks, PathSeq, SRSA, CaPSID, SURPI, PathoScope, VirusScan, MetaShot, ConStrains, RINS, and GRAMMy which have been used or

have potential for such microbial diversity evaluations. These methodologies could help shed light to the role of the microbiota in cancer racial related disparities especially in infection associated-cancers. Further phylogenic and protein functional predictions from bioinformatics pipelines such as PICRUSt, Tax4Fun and ShortBRED among others, provide important clues in understanding microbial differences and commonalities as well as the potential impact on differential outcomes. These tools help us better understand the role of microbes in cancer pathogenesis and identify differences among ethnic groups. Differences in ethnic groups may highlight effectors that impact the treatment decision making process and potential for targeted therapies to help reduce or eliminate cancer disparities across the continuum of cancer disease.

2.2.9 **Literature Cited**

Located in Appendix **E. Literature Cited, complete list**, pp. 147

CHAPTER III. RESULTS

Results section is divided into two manuscripts, "***The landscape of bacterial presence in tumor and adjacent normal tissue across tumor types using raw exome sequencing data from The Cancer Genome Atlas (TCGA) cohorts***," which describes microbial profiles of a subset of 9 cancer cohorts from the TCGA data. The second manuscript is entitled, "***Bacterial diversity correlates with survival in infection-associated cancers of the head & neck, liver and stomach***," that touches on the correlation of microbial co-occurrence in viral associated cancers of the head & neck, liver and gastric which are disproportionately more common among racial minorities.

3.1 **The landscape of bacterial presence in tumor and adjacent normal tissue across tumor types using raw exome sequencing data from The Cancer Genome Atlas (TCGA) cohorts**

3.1.2 **Abstract**

Objective: Several studies have evaluated the viral composition in human tumors using unmapped to human sequencing data, bacterial composition derived from human whole exome sequencing data is less explored. We interrogated bacterial presence in tumor and adjacent normal tissue pairs utilizing whole exome sequencing to generate microbial profiles. Identification of microbial composition directly from tumor tissue permits studying the relationship between microbial changes and cancer pathogenesis. **Methods**: We screened 4777 whole exome sequencing files from 813 cases within 9 TCGA cancer cohorts including carcinomas and adenocarcinomas of stomach (STAD), liver (LIHC), colon (COAD), rectal (READ), lung (LUAD) and bladder (BLCA) as well as squamous cell carcinomas of head and neck (HNSC), lung (LUSC) and cervix (CESC). Data files were processed through a bioinformatics pipeline designed to generate microbial profiles from raw whole exome sequences. Viral DNA presence was used as internal validation and for microbial co-occurrence analyses. Diversity metrics were calculated to compare taxa within sample, between tumor and its paired adjacent normal and across cancer cohorts. Differential abundance was examined using Wilcoxon signed-rank and rank sum tests within each tissue type and between tumor and adjacent normal tissue. Bacterial taxa with false discovery rate (FDR) adjusted p value < 0.05 were considered significantly different at both genus and species level. Presence of *Helicobacter pylori*, *Fusobacterium nucleatum* and *Selenomonas sputigena* was confirmed by quantitative polymerase chain reaction in a cross-sectional fashion with paired tumor and adjacent normal formalin-fixed, paraffin-embedded tissue from an independent population from the Hawaii Tumor Registry. **Results**: Microbial profiles for stomach, liver, colon, rectal, lung, head & neck, cervical and bladder TCGA cohorts were generated. Several taxa were found to be common across cancer types at all classification levels. There was a difference in species richness between tumor and its adjacent normal with significant difference observed in alpha and beta diversity metrics. **Conclusion**: We demonstrate the ability to identify differential composition of bacteria species from human tissue whole exome sequencing data. Paired analyses revealed significant differences in bacterial shifts across STAD, LUSC, COAD and HNSC cohorts whereas little or no differences were evident in CESC, LUAD, LIHC, READ and BLCA cohorts in adjusted models. Profiles are suggestive of microbial shift changes with advanced disease. Experimental validation confirmed 60%-80% of the predicted bacterial differential abundances. Our data highlights the importance of analyzing adjacent tissue which can be indicative of cancer stage progression and the need to examine both, microbial relative abundance and rate of positivity within the sample population.

3.1.3 **Introduction**

Studies evaluating the role of infectious agents in the etiology of cancer have mainly concentrated in the possible role of single organisms. Newer studies have examined the role of the gut microbiota in gastrointestinal and distant cancers (Grat et al. 2016, Routy et al. 2018, Golombos et al. 2018, Gopalakrishnan, Spencer, et al. 2018). Bacteria have been associated with cancer progression exerting beneficial or detrimental effects depending on the time and site of the colonization (Nauts 1989, Schwabe and Jobin 2013). Their highly site specific colonization enables modulation of the tumor microenvironment (Mager 2006, Chang and Parsonnet 2010). Microbial-host dynamics can promote or inhibit host immune response (Paulos et al. 2007, Elinav et al. 2013). These changes lead to the accumulation of insults and epigenetic changes that can change the course of a developing or established tumor(Hattori and Ushijima 2016). Evidence demonstrates that infection-associated cancer subtypes are molecularly distinct (Cancer Genome Atlas Research 2014, Cancer Genome Atlas 2015), which highlights the importance of microbial modulation within the tumor cells. These findings are significant and reveal important microbial patterns and mechanistic pathways in host response to cancer. However, these can be limited in that they do not examine the microbial community composition directly within the tumor and the surrounding tissue microenvironment.

Microbial presence within the tumor and adjacent tissue can inform disease progression, and bacterial roles in cancer pathogenesis (Thomas et al. 2016). Presence information can be derived from human whole exome sequencing data (Tae et al. 2014), similar to transcriptomics or metagenomics methods. Bioinformatics tools facilitate profiling of tumor virome and bacteriome using human sequencing data in the context of cancer-associated pathogenesis (Kostic et al. 2011, Borozan et al. 2012, Naccache et al. 2014, Hong et al. 2014, Chen et al. 2013). However, most studies extract microbial (viral, bacterial and other) using RNA sequencing data. For example, Khoury et al. (2013) interrogated 18 TCGA cancer cohorts sifting through 3775 cases to characterize viral DNA presence and integration sites within tumor tissue derived RNA sequencing data. Here Khoury described important HPV, HBV and HHV-4 differential integration sites across TCGA cancer cohorts and highlighted the utility of RNA sequencing data for tumor virome characterization (Khoury et al. 2013). However, this study lacked validation in tumor tissue either direct or cross-sectional. Similarly, Tang et al. (2013), examined viral gene expression and host fusion, building a viral expression map across 19 of TCGA cancer cohorts using RNA sequencing data (Tang et al. 2013). Salyakina et al. (2013), -9 cohorts, and Cao et al. (2016), -23 cohorts, also examined viral expression in tumor and normal specimens within TCGA cohorts (Salyakina and Tsinoremas 2013, Cao et al. 2016). Cao et al. (2016) demonstrated the ability to identify associations between different viral strains and ethnic differences (Cao et al. 2016). These works were all based in RNA sequencing derived pathogen information. Cantalupo et al. (2018) on the other hand, examined viral integration using RNA, whole exome and whole genome sequencing data across 22 of the TCGA cancer cohorts mapping viral prevalence differences and commonalities within the sample population (Cantalupo, Katz, and Pipas

2018). None of these experimentally validated their findings. Similar to viral profiling, RNA sequencing data has been used for bacterial composition characterization. Riley et al. (2013) examined bacterial DNA integration in 852 TCGA tumor and normal specimens (Riley et al. 2013). They discovered significant bacterial gene integration within various TCGA cohorts. However the highest integration rates were detected in cohorts for which no matched or paired normal sample data was available including stomach adenocarcinoma and acute myeloid leukemia (Riley et al. 2013). Robinson et al. (2017) later examined potential bacterial contamination across 5 TCGA cohorts including acute myeloid leukemia, breast, gliobastoma, ovarian and stomach adenocarcinomas using RNA sequencing data from paired tumor and adjacent normal samples (Robinson et al. 2017). It was found that potential contaminants were present across all cohorts such as *Staphilococcus epidermis*, *Cutibacterium acnes* and *Ralstonia* species after controlling for batch effects (Robinson et al. 2017). Like Riley (2013), Robinson et al. (2017) did not include experimental validation. Additional bacteriome identification has been completed. Zhang et al. (2015) developed a workflow for the identification of low abundant microbial species using whole exome and RNA sequencing data derived from the human genome project based on PathSeq (Zhang et al. 2015). Zhang offered experimental validation in gastric biopsies and TCGA whole genome sequencing data (Huo, Zhang, and Yang 2012). This study demonstrated the ability to identify low abundant microbes in relation to host. Other studies examining tumor microbiota derived from human sequences have looked at mutation interaction and gene expression associations in one or few cancers. Thompson et al. (2017) examined bacteria composition in TCGA breast cancer cohort and associated expression profiles with direct 16SrRNA sequencing validation of bacterial presence with samples from whom RNA sequencing data had originally been derived (Thompson et al. 2017). Thompson found that bacteria presence correlates tumor growth pathway genes. While Greathouse et al. (2018) examined the lung microbiome and the association with TP53 mutation using 16SrRNA can confirm findings with TCGA lung cancer data (Greathouse et al. 2018). Taken together these studies highlight the feasibility of microbial profile identification and functional characterization, however the use of RNA sequencing data may not be the best approach at characterizing bacteria signatures. Use of RNA sequencing could amplify cDNA library artefacts rather than a true reflection of actual RNA abundance and reflects presence of pathogens that are actively being expressed at one specific time (Wang, Gerstein, and Snyder 2009). On the other hand whole exome sequencing data represents the gene expression profiles with over 85% of the known disease causing variants (Choi et al. 2009). Therefore to identify potential association of bacteria to in cancer pathogenesis, whole exome sequencing may provide a better picture. To our knowledge, no works have yet examined cross-cancer microbial composition differential profiling using whole exome sequencing data from tumor and adjacent normal in a strict paired design. We interrogated tumor and adjacent normal tissue from a paired solid cancers cases from the Cancer Genome Atlas, generating bacterial composition across 9 cancer types encompassing 3758 total tumor and adjacent solid tissue

normal samples. To validate results, we performed quantitative polymerase chain reaction (qPCR) in selected differentially abundant taxa with an independent sample population.

3.1.4 **Results**

3.1.4.1 Identification of microbial sequences in TCGA WXS data

We generated microbial profiles for 1690 samples representing 813 cases across 9 human cancers within the TCGA cohorts. We examined raw whole exome sequencing (WXS) files from 441 STAD, 376 LIHC, 443 COAD, 168 READ, 502 LUSC, 582 LUAD, 527 HNSC, 305 CESC, and 412 BLCA. Combined, over 223 billion reads were processed (**Table 6**). Previously described methods (1.5.2 Methods and Planned Statistical Analyses) were used to characterize the bacterial reads derived from human sequences. Briefly, 3758 sequencing files from primary tumor and adjacent normal tissue were pre-processed using SAMtools and Picard to extract unmapped-to-human reads (Li et al. 2009, BroadInstitute 2018). Sequences were then trimmed and quality filtered, and processed using a series of filtering and alignment steps based on a modified PathoScope 2.0 workflow (Hong et al. 2014, Zhang et al. 2015) to subtract additional human sequences and determine microbial relative abundance present within the sample (**Figure 1** Bioinformatics Pipeline). Primary tumor and its paired adjacent normal with detected bacterial reads on either sample were selected at a 1:1 ratio for analyses (**Table 7**). In Table 6 the number of paired cases selected along with the total number of sequencing reads distribution per cancer type is displayed. READ had the largest proportion of reads per sample with 6 billion reads among 36 samples followed by COAD with 28 billion reads. Table 7 displays a comparison of aligned, unmapped, mapped to human, and mapped to microbe reads found within 9 TCGA cancer cohorts. We observed an average microbe to human read ratio of 0.005%. CESC presents the highest ratio of 0.04%. Overall, approximately 99% of the total reads detected in CESC were of viral origin.

Table 6 *Total sample sequencing files screened and processed per TCGA cancer cohort for microbial composition characterization*

TCGA Cohort	Total WXS Files Screened	Total Samples Screened	Total Paired Cases	Total Reads
STAD	443	197	88	2.54E+10
LIHC	376	168	84	2.44E+10
COAD	443	182	88	2.80E+10
READ	168	36	18	6.06E+09
LUSC	502	454	221	6.57E+10
LUAD	582	411	200	5.24E+10
HNSC	527	150	69	1.30E+10
CESC	305	17	8	1.29E+09
BLCA	412	76	37	7.46E+09
Totals	3,758	1,690	813	2.24E+11

WXS=whole exome sequence. Table shows total binary alignment/map (BAM) format raw WXS sequencing files meeting selection criteria and total of paired tumor and solid tissue normal within each cohort screened and selected for bacterial presence.

Table 7 *Sequence comparison of aligned reads in 9 TCGA cancer cohorts*

TCGA Cohort	Human Reads	Microbe Reads	Microbe Reads in Tumor	Microbe Reads in Adjacent
STAD	2.52E+10	3.67E+04	2.85E+04	8.27E+03
LIHC	2.44E+10	2.11E+04	1.52E+04	5.84E+03
COAD	2.78E+10	5.26E+05	1.77E+05	3.49E+05
READ	6.04E+09	2.31E+04	1.03E+04	1.28E+04
LUSC	6.56E+10	6.03E+04	9.62E+03	5.06E+04
LUAD	5.25E+10	1.99E+04	9.18E+03	1.08E+04
HNSC	1.30E+10	7.30E+03	4.40E+03	2.90E+03
CESC	1.28E+09	5.67E+05	5.38E+05§	2.91E+04§
BLCA	7.40E+09	4.40E+03	3.87E+03	5.24E+02
Totals	2.23E+11	1.27E+06	7.95E+05	4.70E+05

Table shows total sequence alignment/map (SAM) format aligned sequencing reads. Human reads is the cumulative total number of mapped to human reads after filtering and subtraction steps. §In CESC the proportion of bacterial reads was 5% and 6% in tumor and adjacent normal respectively

We detected bacterial DNA presence in 94% of the primary tumors and 92% adjacent solid tissue normal samples (**Table 8**). Viral DNA presence, mainly HPV, HBV and EBV, was used as internal pipeline validation and for microbial co-occurrence analyses. Our pipeline was designed to identify DNA sequences as such, RNA viruses like HCV were not detected. Viral DNA presence was detected in 33% and 35% of the tumor and adjacent solid tissue normal samples respectively. The highest proportion of viral DNA positivity was detected in colon and cervical cancers. Colorectal (COAD and READ) and HNSC cohorts were found to have the highest proportion of cases with bacterial DNA. The lowest proportion of samples with any bacterial DNA presence were observed in BLCA cancer cohort (76% of the samples), while the lowest proportion of microbial to human reads ratio were observed in LUAD cohort.

Table 8 *Proportion of samples with microbial reads at any detection level*

TCGA Cohort	Samples with Bacteria in Tumor N (%)	Samples with Bacteria in Adjacent N (%)	Samples with Virus in Tumor N (%)	Samples with Virus in Adjacent N (%)
STAD	74 (84)	73 (83)	30 (34)	35(40)
LIHC	68 (81)	66 (79)	13 (15)	17 (20)
COAD	88 (100)	88 (100)	88 (100)	88 (100)
READ	18 (100)	18 (100)	11 (61)	10 (56)
LUSC	211 (95)	209 (94)	63 (28)	81 (36)
LUAD	200 (100)	193 (97)	41 (21)	34 (17)
HNSC	72 (100)	71(100)	11 (16)	9 (13)
CESC	8 (100)	7 (88)	8 (100)	8 (100)
BLCA	28 (76)	28 (76)	2 (5)	1 (3)
Totals	767 (94)	753 (92)	267 (33)	283 (35)

Included cancer types and number of samples with microbial presence at any level. No difference was observed in the number of samples with microbial presence between tumor and its adjacent solid tissue normal for any cohort.

3.1.4.2 Population Characteristics

We analyzed 813 paired primary tumor and adjacent solid tissue normal cases from TCGA. From these, 85 STAD, 81 LIHC, 88 COAD, 18 READ, 221 LUSC, 200 LUAD, 69 HNSC, 8 CESC, and 28 BLCA with microbial reads were analyzed (**Table 9**). Table 9 displays population demographics and clinical characteristics for all 9 cohorts. Clinical data was available for 746 cases. Only paired samples with available clinical data were used in association analyses. 69% of cases were White (independent of Hispanic origin), 9% were African American (independent of Hispanic origin), 4% Asian, and 1% were of other racial groups. Age at diagnosis ranged from 20 to 90 years of age with a mean of 64 years (SD ±11.9). There were some expected differences in age at diagnosis among the cancer cohorts with the youngest population belonging to the CESC cohort (47±13.5.). There was an 8% difference in the overall proportion of females to males (54% vs 46% respectively). 48% of the tumors were classified as Stage II-III.

3.1.4.3 Taxonomic Composition across Cancer Types

Our pipeline detected 1,264,775 quality WXS microbial reads representing 1353 unique bacteria taxa from which 882 were shared across cancer cohorts (**Table 10**). From these, 12 species were present in all cohorts including *Actinomyces oris*, *Bradyrhizobium sp. BTAi1 Bradyrhizobium sp. ORS, Cutibacterium acnes, Escherichia coli, Leptothix cholodnii, Neisseria sicca, Ralstonia insidiosa, Rhodopseudomonas palustris, Shingomonas melonis, Sphingomonas panacis and Bradyrhizobium diazoefficiens,* and 24 species, *Bacillus subtilis*, *Cutibacterium acnes*, *Escherichia coli*, *Mycoplasma mycoides*, *Corynebacterium pseudotuberculosis*, *Ralstonia pickettii*, *Bacillus mycoides*, *Mitsuaria* sp. 7, *Streptomyces gilvosporeus*, *Bacteroides fragilis*, *Roseateles depolymerans*, *Psychromicrobium lacuslunae*, *Bacteroides thetaiotaomicron*, *Bacteroides dorei*, *Bacteroides ovatus*,, *Bacteroides vulgatus*, *Bacteroides caecimuris*, *Alistipes finegoldii*, *Bradyrhizobium sp. BTAi1*, *Rothia mucilaginosa*, *Flavonifractor plautii*, *Arthrobacter sp. IHBB 11108*, *Sphingomonas koreensis* and *Roseburia hominis* were found to frequently co-occur in tumor and adjacent normal samples across cohorts and are described in more detail in the next section (3.1.6.4 Core Microbiota). Taxonomic composition was found to be similar to that previously reported in RNA-seq, WGS or WXS data (Cantalupo, Katz, and Pipas 2018, Zhang et al. 2015, Yu et al. 2017, Robinson et al. 2017, Khoury et al. 2013). One of the important findings is bacteria shifts observed in tumor compared to adjacent normal in most cancer types. Eight major phyla were found among all cancer cohorts with significant differential relative abundances (pvalue <0.05). Taxa from the Proteobacteria phylum were found in all nine cancer types. Bacteroidetes were highest in COAD. Firmicutes were higher in STAD tumor and BLCA adjacent normal compared to their paired corresponding tumor tissues. Fusobacteria were present at low levels across various cancer types including STAD, LUSC, HNSC and COAD. **Figure 9** displays the landscape of bacterial reads proportions in tumor compared to paired

Table 9 *Demographics and clinical characteristics summary*

		STAD N= 85	LIHC N =81	COAD N=88	READ N=18	LUSC N=221	LUAD N =200 *148	HNSC N=69	CESC N=8	BLCA N=28	Totals N=746*
Race	N (%)										
	White	54 (64)	64 (79)	37 (42)	8 (44)	147 (67)	120 (81)	56 (81)	4 (50)	25 (89)	515 (69)
	African American	3 (3)	6 (7)	7 (8)	1 (6)	16 (7)	23 (16)	9 (13)	1 (13)	2 (7)	68 (9)
	Asian	16 (19)	7 (9)	--	--	3 (1)	2 (1)	1 (1)	--	--	29 (4)
	Other Race	--	--	1 (1)	--	--	1 (1)	--	2 (25)	--	4 (1)
	Not reported	12 (14)	4 (5)	43 (49)	9 (50)	55 (25)	2 (1)	3 (4)	1 (13)	1 (4)	130 (17)
Age at diagnosis											
	Mean, ±SD,	67 ±10.5	64 ±14.7	71 ±12.3	63 ±14.6	68, ±8.4,	65 ±10.3	63 ±12.2	47 ±13.5	69 ±10.7	64 ±11.9
	Range	41-88	20-86	40-90	40-90	40-85	40-87	26-88	22-69	48-90	20-90
Sex	N (%)										
	Male	48 (56)	46 (57)	47 (53)	7 (39)	64 (29)	63 (43)	48 (70)	--	19 (68)	342 (46)
	female	37 (44)	35 (43)	41 (47)	11 (61)	157 (71)	85 (57)	21 (30)	8 (100)	9 (32)	404 (54)
Tumor Stage	N (%)										
	I	14 (17)	33 (41)	11 (13)	4 (22)	121 (55)	79 (53)	1 (1)	4 (50)	3 (11)	270 (36)
	II-III	47 (55)	35 (43)	62 (70)	9 (50)	96 (43)	61 (41)	30 (43)	4 (50)	11 (39)	355 (48)
	IV	8 (9)	3 (4)	14 (16)	4 (22)	3 (1)	6 (4)	38 (55)	--	14 (50)	90 (12)
	No staging	16 (19)	10 (12)	1 (1)	1 (6)	1 (1)	2 (1)	--	--	--	31 (4)

Basic demographics characteristics from cases with available clinical data. *Fraction with clinical data; -- not applicable/no data; Other Race includes groups with less than 1 sample from Native American, Alaska Native, Native Hawaiian or other Pacific Islanders background described as other to maintain privacy.

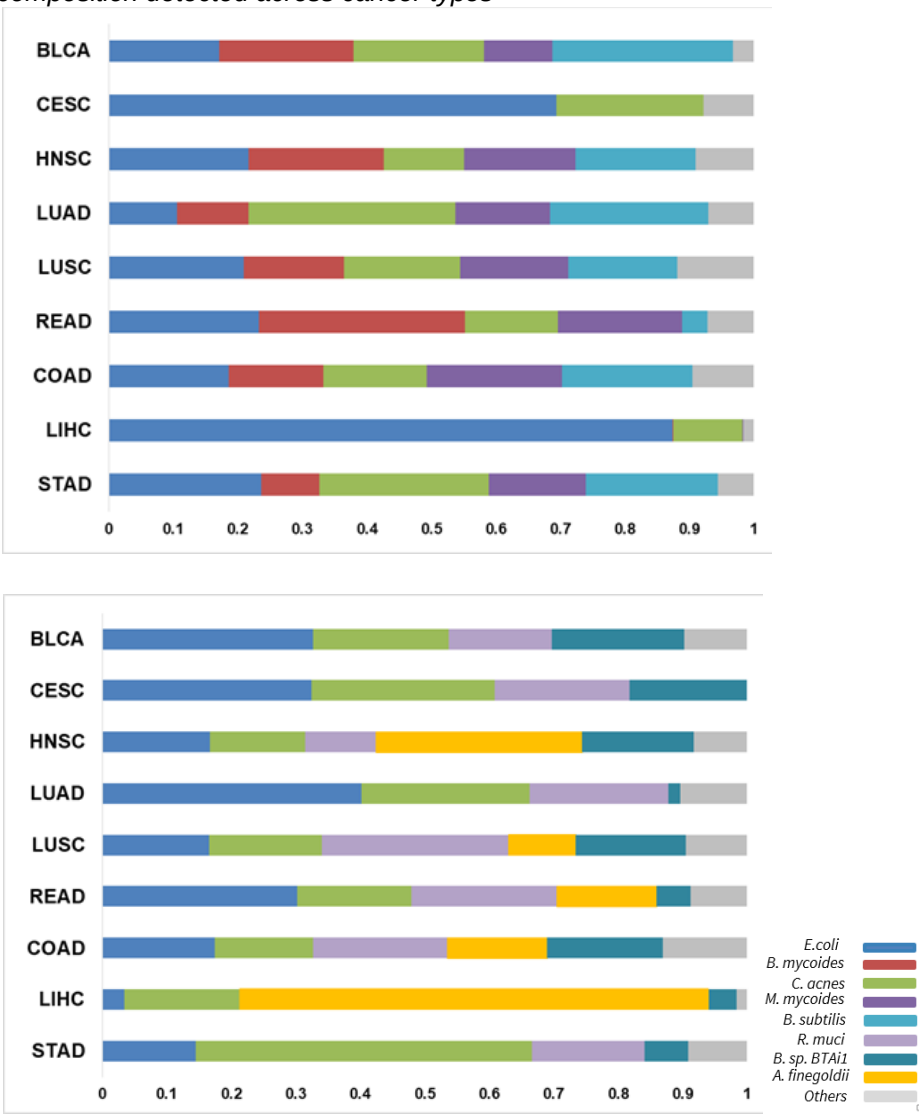
adjacent normal with greater that 1% at the phylum level are shown. Shifts in bacteria abundance were evident in STAD, COAD, CESC and BLCA cancers while less evident in LIHC, LUAD and READ, where the bacterial composition between tumor and its adjacent normal were virtually indistinguishable at the phylum level. STAD had a significant increase in Bacteroidetes (6% in adjacent normal compared to 11% in tumor) and Firmicutes which, composed nearly half of the total reads found in tumor while less than 10% in adjacent normal (Figure 9-B). A 15% decrease was observed in the Proteobacteria like species in STAD. In COAD, we detected a significant decrease of Bacteroidetes (-21%) in tumor, while Proteobacteria (+20%) levels were increased compared to adjacent solid tissue normal. In LUSC, there was a slight opposite shift between Proteobacteria (lower) and Actinobacteria (higher) in tumor compared to adjacent solid tissue normal. In addition, the tumor appeared to be colonized by higher levels of Firmicutes which were relatively absent in the adjacent normal tissue. In HNSC, Actinobacteria decreased by 17%, while Bacteriodetes increased 9% in tumor compared to adjacent normal. CESC had a significant change in the number of Actinobacteria colonizing the tumor tissue, almost entirely shifting to 1% compared to 25% in adjacent normal. BLCA cohort cancer appeared to have the greatest shift change in composition where tumor was colonized almost entirely by Proteobacteria (98% vs to 50%) compared to adjacent normal. Three species, *Escherichia coli, Cutibacterium acnes* and *Bradyrhizobium sp. BTAi1* were found to be present in all 9 cohorts. Several *Bradyrhizobium* spp. and *Escherichia* spp. strains were detected in multiple cohorts at different rates in either tumor or adjacent normal. *Escherichia coli* and *Cutibacterium acnes*, were found to be consistently present across samples. Other species were found across all cohorts with at least 1 read, including those before mentioned and *Bradyrhizobium* sp. ORS 285, *Leptothrix cholodnii*, *Nisseria sicca*, *Rasltonia insidiosa*, *Rhodopseudomonas palustris*, *Sphingomonas melonis* and *Sphingomonas panacis*. ***Figure 7.*** Measures of total read absolute abundance, reads proportional relative abundance and percent prevalence provided different interpretation regarding the taxonomy compositional structure. We point out that all three measures must be used for accurate characterization of the tumor microbiota with greater weight to percent population prevalence and relative abundance when identifying clinically relevant taxonomy to avoid erroneous conclusions.

3.1.4.4 Core taxa characterization

Identification of core microbiota is important to the understanding of the tumor microenvironment and the role bacteria play in cancer pathogenesis (Sun and Kato 2016). We therefore wanted to identify differences and commonalities of species shared within each cohort's tumor and adjacent normal pairs and across cancer cohorts in order to identify clinically relevant differences and commonalities in species. A positivity detection threshold of ≥ 0.2% relative abundance was set. Assuming that each identified taxa is present at least once in each sample with a minimum of 1 read, a positivity detection rate of 0.2% was deemed

reasonable. Core taxa was defined as that identified at a minimum positive detection rate, present in the majority of population, and shared between tumor and adjacent normal pairs with a minimum of 20% prevalence in each sample type. Core taxa was calculated based on study assumptions and verified using microbiome-R package (version 1.3.3). Microbial core composition using microbiome package was calculated using default settings and taxa positivity detection rate of 0.2% and prevalence of 20%. All taxa identified for each cohort were evaluated in all cohorts regardless if met core criteria.

Figure 7 *Core taxa composition detected across cancer types*

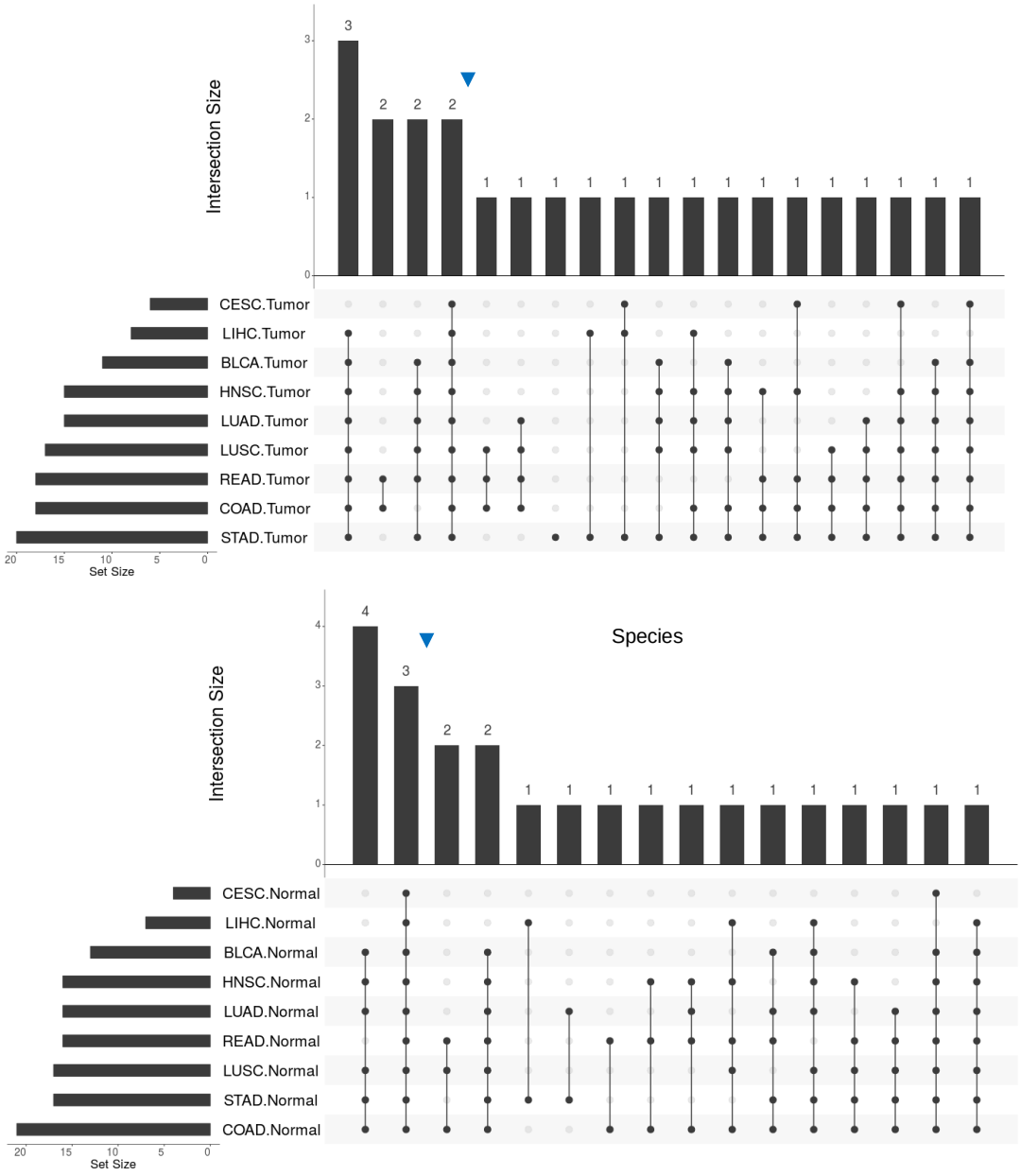


Mean relative abundance (normalized proportions) of top 10 bacteria species most commonly identified within each cohort and defined as core taxa in tumor (L) and adjacent normal (R).

Top core taxonomic identification is shown in **Figure 7**. *Bacillus subtilis* was the most frequent taxa identified in the population, yet the based on proportion of reads and relative abundances across cohorts

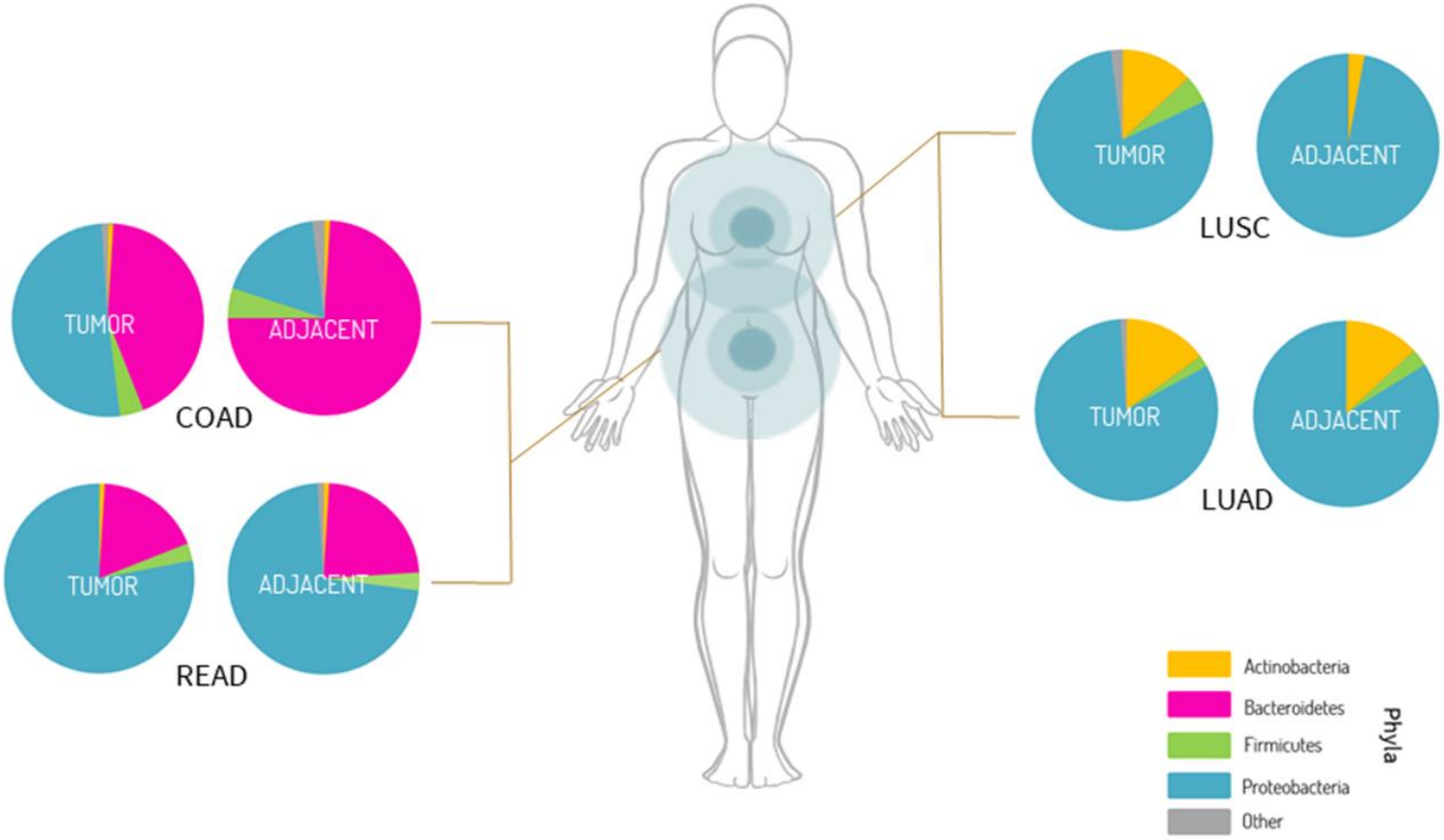
was very low and its presence is masked by higher read taxa. COAD had the highest proportion of taxa present across all cohorts with 467 species shared between tumor and adjacent normal. It also had the

Figure 8 *Core taxa shared species across cancer types*



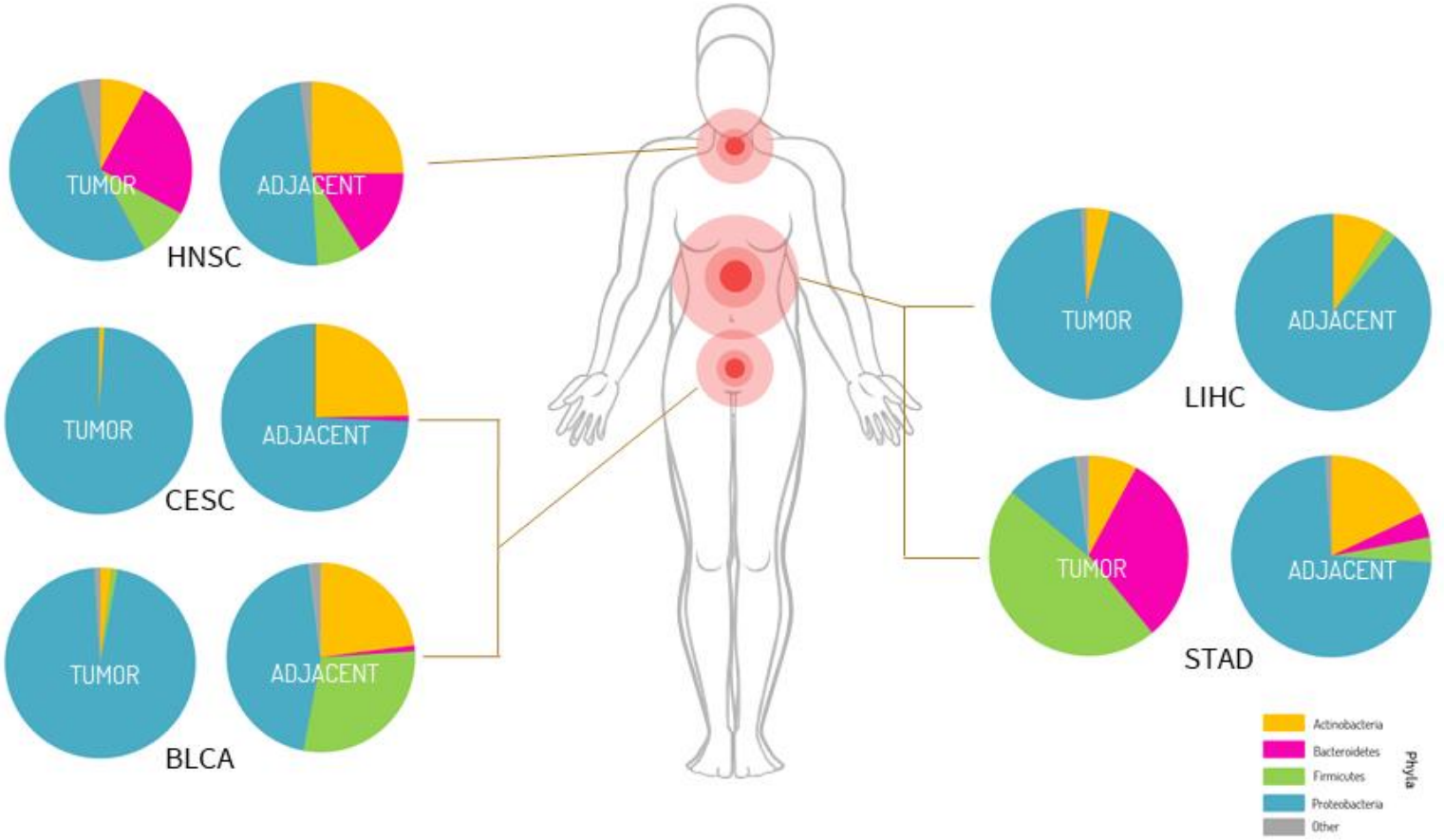
Compositional bar graph showing size of individual core taxonomies (left horizontal bars) and intersect of shared species in tumor and normal samples across cohorts. 24 species were found to frequently co-occur in tumor and adjacent normal samples across cohorts. Each column represents intersect of shared species. In tumor samples 2 species are shared across all 9 cohorts, *Escherichia coli* and *Cutibacterium acnes* (blue arrow top), and 3 species are shared in all 9 cohorts in adjacent normal, *Escherichia coli* and *Cutibacterium acnes,* and *Bradyrhizobium sp. BTAi1* (blue arrow bottom). Other species are shared across cohorts at different rates

***Figure 9**-A Landscape of bacterial shift changes in tumor and adjacent normal tissue across tumor types*



Proportion of bacterial reads and composition shift in tumor and adjacent normal paired samples at the phylum level illustrated with pie charts. Figure shows anatomical site proximity between cancer types. Phyla >1% is shown. Reads proportions in tumor compared to adjacent normal in READ and LUAD cohorts are indistinguishable at phylum level while greater shifts are observe in COAD Bacteroidetes and LUSC Actinobacteria tumor to adjacent normal ratios.

***Figure 9**-B Landscape of bacterial shift changes in tumor and adjacent normal tissue across tumor types*



Proportion of bacterial reads and composition shift in commonly infection-associated cancers tumor and adjacent normal paired samples at the phylum level illustrated with pie charts. Phyla >1% is shown. Figure shows anatomical site proximities between cancer types. Proportion of Actinobacteria, Firmicutes and Proteobacteria are observed for most infection –associated cancers

Table 10 *Taxonomy classification counts distribution across cancer cohorts per sample type*

Cancer Cohort		Phylum		Class		Order		Family		Genus		Species		Mean Difference
		Tumor	Adjacent	Tumor	Adjacent	Tumor	Adjacent	Tumor	Adjacent	Tumor	Adjacent	Tumor	Adjacent	
STAD	(n=706)	14	9	30	21	62	58	125	86	246	175	558	419	
	Shared taxonomy		9		19		58		76		136		268	
											α	24.4	18.6	5.8
											β	31.6	31.9	-0.3
LIHC	(n=290)	6	4	10	8	24	22	45	46	80	89	203	190	
	Shared taxonomy		4		6		15		30		47		103	
											α	14.2	12.7	1.5
											β	21.2	21.1	0.1
COAD	(n=1028)	15	17	33	40	72	87	150	186	263	387	663	832	
	Shared taxonomy		13		31		62		138		251		467	
											α	35.5	47.8	-12.3
											β	17.7	16.4	1.3
READ	(n=279)	7	7	19	18	35	36	61	73	99	122	167	223	
	Shared taxonomy		6		16		30		50		75		111	
											α	20.8	29.2	-8.4
											β	7	6.6	0.4
LUSC	(n=741)	7	12	16	26	41	61	84	122	182	281	415	661	
	Shared taxonomy		17		14		37		73		153		335	
											α	10.3	14.9	-4.6
											β	39.2	43.3	-4.1
LUAD	(n=820)	11	11	24	25	56	58	115	112	249	242	639	630	
	Shared taxonomy		10		22		49		96		185		449	
											α	20.3	18.5	1.8
											β	40.2	42.6	-2.4
HNSC	(n=452)	13	8	27	19	47	41	91	88	158	157	340	331	
	Shared taxonomy		8		18		33		71		120		219	
											α	15.3	13.6	1.7
											β	21.4	22.4	-1
CESC	(n=120)	4	3	10	5	18	12	31	15	53	19	103	40	
	Shared taxonomy		3		5		8		7		7		23	
											α	15.8	6.5	9.3
											β	5.5	5.2	0.3
BLCA	(n=195)	5	7	9	12	26	28	46	45	19	63	132	113	
	Shared taxonomy		5		9		23		31		37		50	
											α	5.9	5.5	0.4
											β	14.3	17.1	-2.8

Table displays primary tumor, adjacent normal and shared taxa counts with mean per cohort alpha and beta diversity measures. n=total unique taxa counts per cohort. Boxed represents α (alpha diversity means) and β (beta diversity means defined as a ratio of unique and shared species within each cohort. Yellow bars represent proportion of shared taxa at different taxonomy levels. Significant mean differences in alpha and beta diversity represented with directional bars.

most varied community with a mean alpha diversity difference of -12.3 indicating greater community abundance in the adjacent normal samples (**Figure 10**). In COAD, 12 species were identified as core microbiota, confirmation with microbiome R package resulted in 24 species. List of core taxa is available in Supplemental **Table A1** (pp. 114). We found an inverse proportion of the *Bradyrhizobium* like reads in CESC and LUSC. In CESC the number of *Bradyrhizobium* like reads were 22 times higher in tumor than in its paired adjacent normal, while number in LUSC were higher in the adjacent normal. *Bradyrhizobium* like reads proportion were also significantly higher in COAD tumor than that of its adjacent normal. Like *Bradyrhizobium*, *Escherichia* spp. reads including *Escherichia coli, Escherichia fergusonii* and *Escherichia albertii* were detected in multiple cohorts. *Escherichia* species were inversely proportional in colon and liver samples with a higher abundance in LICH tumors and colon adjacent normal samples compared to their respective pairs. Although Escherichia are present in all cohorts their normalized reads abundance are negligible when compared across cohort.

3.1.4.5 Diversity Metrics, Alpha and Beta diversity

Microbial diversity is associated with cancer and treatment outcomes (Gopalakrishnan, Spencer, et al. 2018). We therefore examined within sample diversity (alpha diversity as defined by Shannon-Weiner), measures of evenness and species richness and between sample diversity (beta diversity) in tumor and adjacent normal tissue paired cases within each cancer cohort. Diversity measures were calculated using vegan package (version 2.5-3) and Microsoft Excel (v.2013). Measures were calculated at the taxonomy or operational taxonomic units (OTU) level and collated at the species level (by aggregating strains and subspecies of the same species). Species richness, the number of species per sample, was overall slightly higher in adjacent normal compared to tumor samples (average total tumor species 358 vs 382 in adjacent normal) Tumor richness was higher among STAD, LIHC, CESC and BLCA cohorts. In COAD, LUSC and READ richness was higher in adjacent normal, while there were no differences noted in in HNSC and LUAD cohorts. Diversity varied by age, sex and histopathological staging at varying degrees across different cancer cohorts and is presented in detail in cancer specific findings.

3.1.4.6 Cancer Specific Findings

We compared the bacterial relative abundances and bacterial diversity in tumor and its paired adjacent solid tissue normal for each cancer type and across cancer groups. All specimens with bacterial reads were considered (**Table 8**). Likewise, per study assumptions all bacterial taxa identified by the bioinformatics pipeline were considered equally important and with equal weight. All taxa were analyzed for differential abundance (**Table 11**). Association testing of each significant taxa with clinical and demographic factors are presented. Clinical relevance was established by relative abundance difference of 15% between samples types.

Table 11 *Fraction of taxa with presence at any detection level for each cancer type*

TCGA Cohort	Total	Fraction of species ≤1 read	Fraction of species >1 ≤10 read	Fraction of species >10 read
STAD	706	283	287	136
LIHC	290	141	110	39
COAD	1028	371	428	229
READ	279	130	98	33
LUSC	741	228	289	224
LUAD	820	259	382	180
HNSC	452	164	189	99
CESC	120	58	37	25
BLCA	195	101	67	27
Totals (%)	4631	1735 (38)	1887 (41)	992 (21)

Table shows Fraction of taxa collated at species level with at least 1 hit. 38% of the identified taxa hits had ≤1 read across all paired samples within each cohort. Read fractions of less than 1 read occurs when sequences map to multiple top hits.

3.1.4.6.1 Stomach

We examined 170 STAD paired primary tumor and adjacent normal sample sequences from 85 cases. From these, 56% were male, average age at diagnosis was 67 years (±10.5 SD). Most were White (63%) and most classified as tumor stage II-III (55%). **Table 9**. Average read per sample was 360 in tumor and 107 in adjacent normal. Average number of species per sample in tumor were 24 compared to 19 in adjacent normal. We detected 1050 bacteria OTUs corresponding to 14 phyla, 32 classes, 62 orders, 135 family, 315 genus and 706 unique bacterial species. There was a significant difference in the proportions of taxa in tumor compared to taxa numbers in adjacent normal (Fisher, p=0.007). Four species, *Bacillus subtilis*, *Arthrobacter* sp. IHBB11108, *Cutibacterium acnes*, and *Mycoplasma mycoides* were found to be present in 20% or more of either sample type. Relative abundance of *Selenomonas* species were consistently higher in tumor compared to adjacent normal across samples. *Selenomonas sputigena* was the most prevalent species in tumor samples with 13% of the total reads in tumor (Figure 10), while *Helicobacter pylori* strains made 60% of the total reads in adjacent normal. Wilcoxon signed rank test was used to examine the differential abundance between tumor and adjacent normal pairs. Differential relative abundance for 10 taxa representing 4 major phyla and 7 genus levels were found to be higher in tumor than in adjacent normal (Figure 11). However this difference was not statistically significant (p=<0.05 FDR= 1). Presence of *Helicobacter pylori*, was found to be significantly higher in the adjacent normal compared to tumor tissue (log2fc = -4.8, p=<0.0001 FDR= 0.01).

Because gastric cancers molecular subtypes are associated with Epstein Barr virus, we evaluated presence of HHV-4 as internal validation and co-infection analyses. HHV-4 was detected in 25 tumor and 25 adjacent normal samples. Status of HHV-4 did not differed within the paired sample population.

However, the proportion of reads detected in tumor were significantly higher than those detected in adjacent normal tissue samples at a ratio of 102:1. There was no correlation between HHV-4 status and *Helicobacter pylori* presence however there was a positive trend of *Helicobacter pylori* and HHV-4 infection status (correlation coefficient=0.17, 95%CI=0.02-0.31, p=0.03). Presence of *Helicobacter pylori* was positively correlated with histopathological type. In males, *Helicobacter pylori* status was strongly correlated with intestinal type papillary adenocarcinoma (likelihood ratio= 5.5, p=0.02, phi=1). In females there was a weak correlation between *Helicobacter pylori* and unspecified type carcinoma (likelihood ratio 4.8, p=0.03, phi=0.32). *Fusobacterium nucleatum* was detected in 9% (n=8) with a median relative abundance of 0.001 (range 0.002, 0.25). *Veillonella parvula* was 4 times higher in tumor compared to adjacent normal (log2fc=4.5, p=0.03, FDR=1) though not significant after multiple test correction. The odds of tumor presence of *Vellionella parvula* in tumor were 3 times the odds of presence in adjacent normal (OR: 3.2, 95%CI: 1.1-9.2, p=0.03). In tumor tissue presence of *Selenomonas sputigena* was correlated with presence of *Fusobacterium nucleatum* (correlation coefficient=0.48, 95%CI 0.30, 0.63, p=<0.001) and weakly correlated with HHV-4 status (rho=0.24) however not with presence of *Helicobacter pylori.* In adjacent normal, presence of *Selenomonas sputigena* was correlated with HHV-4 status and *Helicobacter pylori* presence (rho= 0.31 and rho=0.26 respectively).

Figure 10 *Microbial composition differences in tumor and adjacent normal paired samples for STAD*

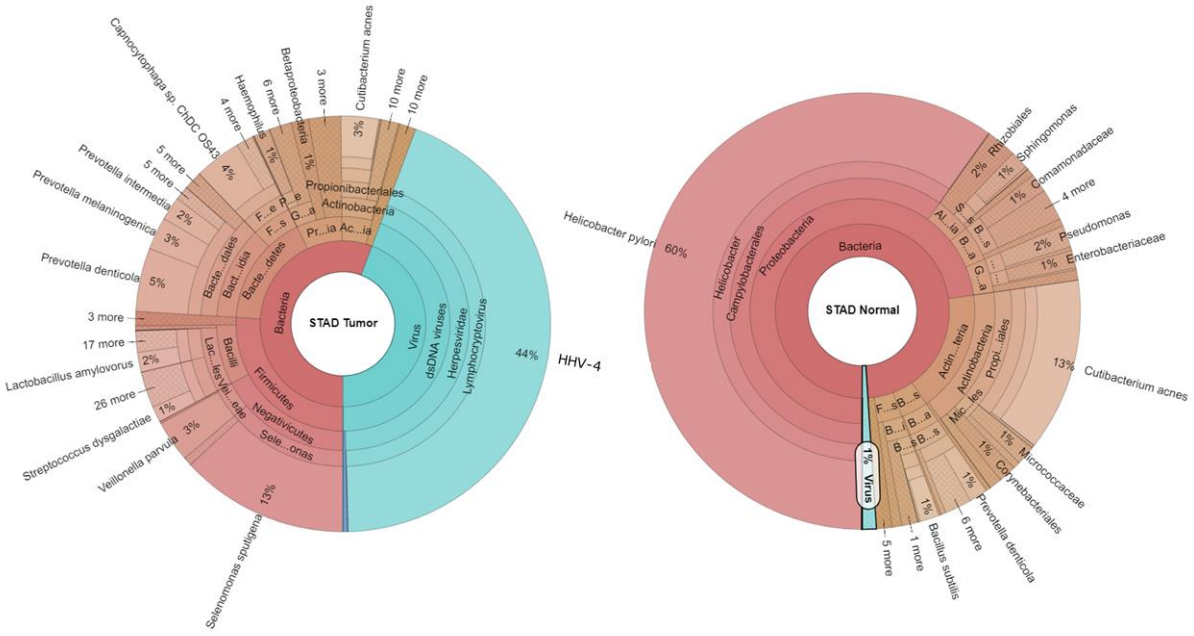


Figure shows Krona plot with compositional differences among 7 taxonomic levels in STAD cohort. Krona plot is colored from red to green clockwise with the most abundant taxa shown in red. Percent within wedges represent relative abundance by total number of reads present in the population. All taxa were included in the data, for graphical purposes, plot shows taxa > 1% abundance.

Figure 11 *Log 2 Fold Change (l2fc) of top taxa in STAD cohort*

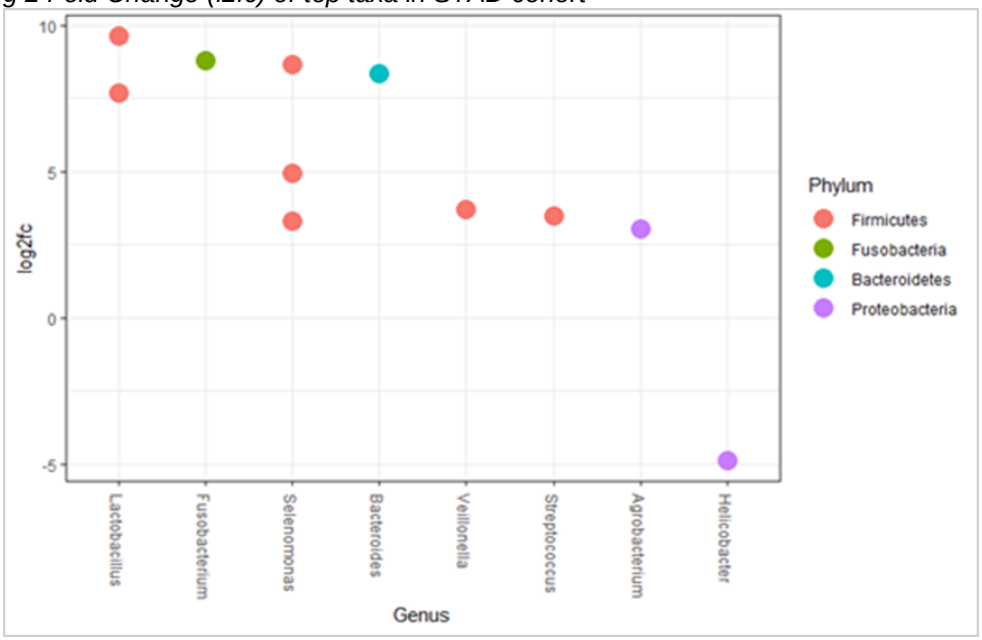
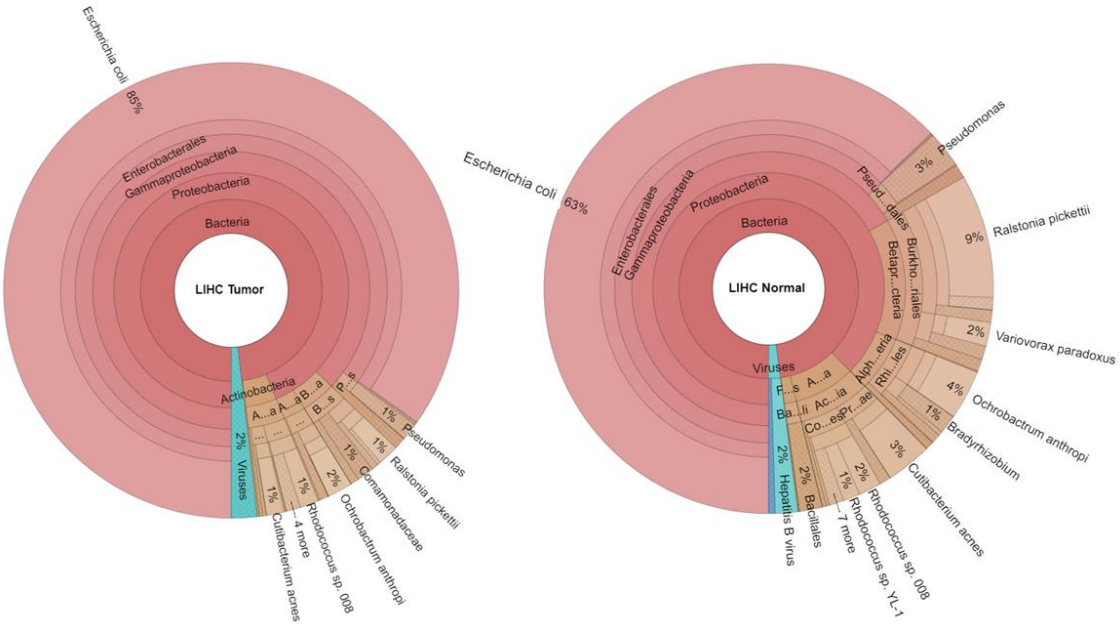


Figure shows log 2 fold change of differential abundant taxa in a paired test in STAD. Each dot represent a species within the genus, colored by phylum. Each species with Wilcox signed rank test p<0.05 are shown. Firmicutes were the most predominant phyla with 7 species. Only *H pylori* (in adjacent normal) was statistically significant after adjustment for multiple testing (FDR).

3.1.4.6.2 Liver

A total of 162 samples from 81 paired cases were examined. The majority of the cases were male (57%), White (79%), and mean age at diagnosis was 64 years (±14.7 SD). In LIHC, 68 out of 81 paired cases were positive for bacterial reads at any detection level. Within these cases, we identified 6 phyla, 12 classes, 31, order, 61, family, 122 genus and 290 unique bacterial species (**Table 10**). There was no difference in the number of taxa present in tumor compared to normal at the OTU level or at collated species level (p=0.62 at OTU, p=0.71 collated at species). Overall, *Escherichia coli* was the most abundant species, detected in 68% of cases (**Figure 12**). We found no difference in bacterial composition between tumor and adjacent normal in paired tests. We note a small 5% decrease in the level of Betaproteobacteria like reads in tumor compared to adjacent normal with *Ralstonia pickettii* the most predominant in both. *Rhodococcus erythropolis* like reads were uniquely identified in adjacent normal tissue as significantly different within the tissue type (p=0.002, FDR=0.05). *Rhodococcus* spp. are known laboratory contaminants. In tumor, *Delftia acidovorans* like reads were also uniquely identified, however difference was not significant after FDR multiple test correction (p=0.006, FDR=0.07). For verification, differential analysis was repeated using count data with DESeq2 package. No differentially abundant taxa was identified by DESeq2.

Figure 12 *Microbial composition in tumor and adjacent normal paired samples for LIHC*



Krona plot with compositional differences among 7 taxonomic levels in LIHC cohort. Proteobacteria dominated both tumor and adjacent normal, with *Escherichia coli* like reads making up 85% of tumor and 63% of adjacent normal total reads. Viral like reads represented 2% of the tumor and adjacent normal (dominated by HBV). All taxa ≥ 1% are shown. Firmicutes made up 2% of adjacent and were relatively absent in the tumor (0.09%) while Actinobacteria made up 9% of the adjacent normal reads compared to 4% in tumor. *Cutibacterium acnes* dominated the adjacent tissue while *Rhodococcus* spp. dominated the Actinobacteria in tumor

Given LIHC viral infectious etiology, HBV and HHV-4 viral reads were evaluated as internal validation and co-occurrence with bacteria presence. These were detected in 10 LIHC cases. The number of HBV reads in tumor were twice the number of reads in adjacent normal. HHV-4, was noted to be present only in tumor samples. HBV and HHV-4 like reads represented <2% of the total tumor or adjacent normal reads. Most samples with positive identification of HBV or HHV-4 did not have bacterial content. Of those with bacterial reads, most commonly co-occurred with Actinobacteria and Proteobacteria like species. Diversity metrics were completed with paired cases with at least 1 bacterial read (**Table 12**). Richness varied by tumor stage and sex, higher males classified as stage I ($p=0.03$) while no differences were observed in alpha diversity when stratifying by tumor stage and sex. There were no differences in beta diversity between tumor and its adjacent normal.

Table 12 *Richness and Diversity in Hepatocellular carcinoma*

		Female		Male		
		Primary Tumor n=35	Adjacent Normal n=35	Primary Tumor n=46	Adjacent Normal n=46	**Pvalue**
Richness-Tumor Stage						
Mean (SD)						
Stage I	(n=66)	12.9 (8.9)	12.3 (8.4)	**20.8** (18.0)	17.8 (14.5)	
Stage II	(n=30)	8.9 (5.6)	12.0 (9.4)	15.2 (16.7)	16.8 (13.4)	
Stage III	(n=42)	11.6 (8.9)	11.3(6.3)	11.6 (9.8)	13.0 (9.1)	
Stage IV	(n=4)	5.0 (0)	10.0 (0)	0 (NA)	1 (NA)	
Not Reported	(n=20)	8.5 (7.5)	1.7 (0.5)	15.5 (7.4)	5.6 (4.3)	**0.034**
Diversity-Tumor Stage						
Mean (SD)						
Stage I	(n=66)	1.65 (1.08)	1.51(0.91)	1.69 (1.06)	1.64 (0.96)	
Stage II	(n=30)	1.38 (0.89)	1.44 (1.08)	1.39 (1.06)	1.46 (1.18)	
Stage III	(n=42)	1.58 (0.91)	1.58 (0.88)	1.21(0.94)	1.10 (0.95)	
Stage IV	(n=4)	1.49 (0)	1.95 (0)	0 (NA)	0 (NA)	
Not Reported	(n=20)	0.11(0.11)	0.46 (0.32)	1.72 (1.06)	0.75 (0.70)	0.415

Table 3.8 displays richness (total number of species) and diversity (Shannon-Wiener index of diversity) by sex and sample type across tumor stage. Overall, males are associated with higher species richness (Estimate=4.5±2.1, p=0.034). In tumor stage I, males have higher species richness compared to their paired adjacent normal (p=0.02), yet there is no difference in species diversity in tumor compared to adjacent normal. Significant numbers in bold.

3.1.4.6.3 Colorectal cancers

A total of 212 samples from 88 colon (COAD) and 18 rectal (READ) paired cases were examined. Most COAD cases did not report race or ethnic backgrounds (49%), 42% were White. Mean age at diagnosis was 71 years (±12.3SD), and 70% were classified as stage II-III (**Table 9**). READ were younger on average with mean age at diagnosis of 63 years (±14.6 SD) and 61% were males. Similar to COAD, 44% were white, while 50% did not report racial background and most were classified as stage II-III. In COAD, average read per sample mapping to bacterial genome was 2006 reads in tumor compared to 3962 reads in adjacent normal. Mean species per sample was significantly different in tumor compared to adjacent normal (mean difference=12.3, p=0.016, 95%CI 2.4, 22.2). **Table 10**. A total of 1028 species corresponding to 19 phyla, 42 classes, 97 orders, 198 families, 433 genus were identified in COAD. For READ average read number per sample was 569 reads in tumor and 708 reads in adjacent normal, while the average number of species per sample was significantly lower in tumor with 20 species per sample compared to 29 in adjacent normal (p=0.008). A total of 420 taxa were identified in READ, from which 279 were unique belonging to 8 phyla, 21 class, 41 order, 84 family, and 146 distinct genus (**Table 10**). Both colon and rectal cancers had a small proportion (<1%) of reads mapping to viral genomes across 7 families. A number of Torque-teno-virus like reads were also identified. **Figure 13.**

Figure 13 *Microbial composition in tumor and adjacent normal paired samples for COAD*

Krona plot with compositional differences among 7 taxonomic levels in COAD cohort. In adjacent normal, 74% of the reads were identified as Bacteroidetes, with a read distribution among several species with the majority identified as B. vulgatus. Proteobacteria reads made up 18% of the adjacent normal reads dominated by E. coli. A significant shift in the proportion of Bacteroidetes to Proteobacteria was observed between tumor and adjacent normal. Proteobacteria increased to 51 % of tumor reads in relation to adjacent normal and an equal reduction in Bacteroidetes (43%) dominated by Bacteroides fragilis. Viral like reads represented <0.1% of both tumor and adjacent normal (dominated by HBV). All taxa ≥ 1% are shown.

In colon samples, overall a significant proportion (66%) of the reads mapped to Bacteroidetes, 28% to Proteobacteria, 4% to Firmicutes, 1% to Verrucombia and 1% to Actinobacteria phyla like reads. Bacteroides to Firmicutes ratio was similar in both tumor and adjacent normal. While a significant shift in the level of Proteobacteria to Bacteriodetes were observed between tumor and adjacent normal (**Figure 9**). The ratio of Proteobacteria to Bacteriodetes was significantly increased in colon tumor compared to its adjacent normal (log2 P/B tumor =0.24, log2 P/B adjacent normal=-2.03). There was no difference in within sample diversity index or the evenness spread (t=1.35, p=0.18, 95%CI=-0.017, 0.005) by sample type. We wanted to know if differences existed when stratifying by sex, age at diagnosis, race and tumor stage. Intra-sample diversity measured by Shannon-Wiener diversity index did not differ by sex or age group (p=0.46). Observable differences by race, tumor stage and site of resection were found. In linear regression model, alpha diversity index differences were not significant after controlling for other variables. Differential abundance were measured by Wilcoxon Signed Rank test. *Bacteroides vulgatus* was found to be significantly different (p=<0.00001, FDR=0.001) between sample types, however the

log 2 fold change was negligible at 0.8 higher in adjacent normal compared to tumor (**Figure 14**). edgeR differential abundance test resulted in 52 species. Similar to Wilcoxon Signed Rank test results, *Bacteroides vulgatus* was identified by edgeR as significantly different with a log fold change of almost 4 times higher in adjacent normal than in tumor (p=<0.00001, FDR=<0.00001, LFC= 3.6). Multiple studies have reported overabundance of *Fusobacterium nucleatum* in tumor tissue associated with colorectal cancer pathogenesis (Castellarin et al. 2012, Kostic et al. 2012, Kostic et al. 2013, Warren et al. 2013, Kumar et al. 2016). Based on these reports we wanted to evaluate the presence of *Fusobacterium nucleatum* in our data set. Overall, Fusobacterial reads represented less than 1% of the total reads mapped in COAD. Of these, 84% were identified as *Fusobacterium nucleatum*. We found that there were considerable differences between detected Fusobacterium reads in tumor and those detected in adjacent normal specimens within the COAD cohort (**Figure 3.6**). The relative abundance means within tumor and within adjacent normal samples differed significantly (tumor p=<0.0001 FDR=0.002, adjacent normal p=0.006 FDR=0.05, respectively). In paired test by Wilcox Sign Rank, mean relative abundance differences were non- significant when comparing tumor to adjacent normal samples (p=0.004, FDR=0.5, l2fc= -1.36).

Figure 14 *COAD Log 2 fold change of significant taxa*

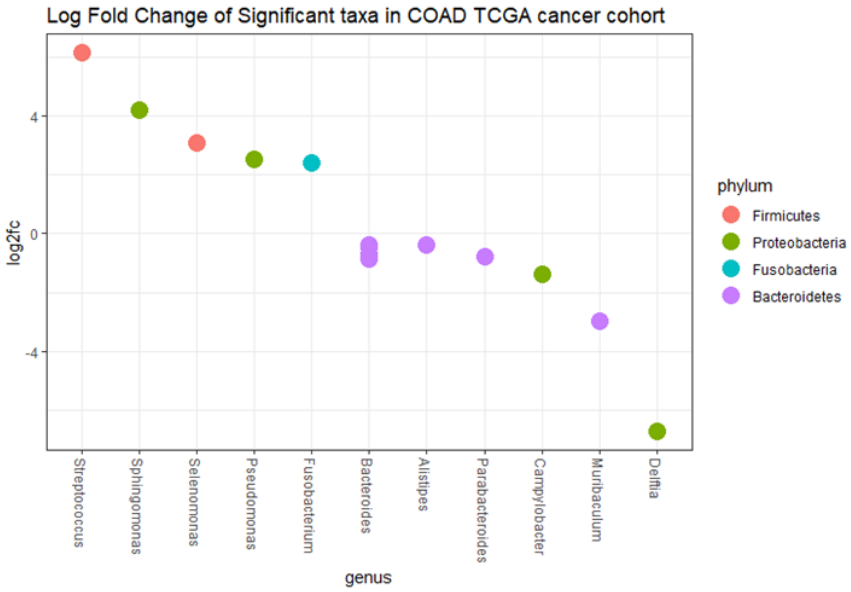


Figure shows fold change (Log2) differential abundant taxa collated at species name in COAD. Each dot represent a species within the genus, colored by phylum. Species above zero are higher in tumor, below are higher in adjacent normal. Bacteroidetes were the most predominant phyla with 5 species. B. vulgatus (in adjacent normal) was statistically significant after FDR correction

Figure 15 *Fusobacteria abundance in COAD cohort*

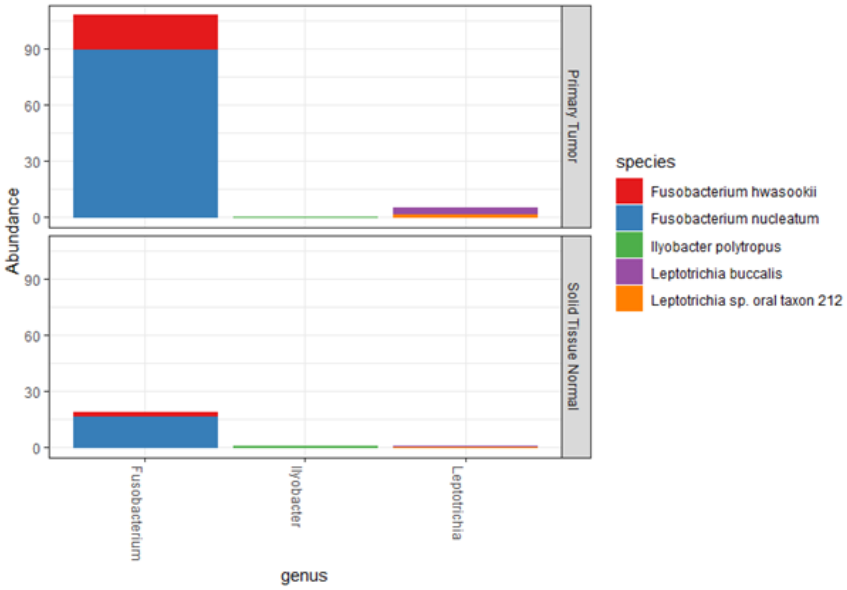


Figure shows Absolute abundance (total reads) of Fusobacteria phyla in COAD by sample type. Each bar represent a genus, colored by species. *Fusobacterium nucleatum* and *Fusobacterium hwasookii* were more abundant in tumor compared to adjacent normal.

In rectal cancer samples there was no difference in the proportions of taxa in tumor compared to taxa numbers in adjacent normal (Fisher, p=0.72). Four Bacteroides species and Escherichia coli were found to be present in 30% or more of either sample type. *Escherichia coli* like reads were the most abundant detected in 86% of the cases. Differential relative abundance did not yield significant taxa. Relative abundance of 3 taxa were consistently higher in adjacent normal compared to tumor while Bacteroides fragilis was 2 times higher in tumor. However this difference was not statistically significant after adjustment for multiple testing (p=<0.02 FDR= 1, l2fc= 1.06). We note a small change in the Proteobacteria to Bacteroidetes ratio, mainly driven by an increase in Enterobacteria (11% increase), decrease in Betaproteobacteria (6%) in tumor and a significant increase of *Bacteroides vulgatus* in adjacent normal. HHV-4 and HPV viral reads were also detected in a small fraction of the samples. HHV-4 were detected in tumor samples and HPV in adjacent normal (Krona plot). There was significant difference in richness by sex and sample type (p=0.008). Species richness was significantly higher in adjacent normal tissue of females compared to males. There were no differences in Shannon alpha diversity or evenness spread by sex (**Table 13**).

Table 13 *Richness and Diversity in rectal adenocarcinoma*

	Primary Tumor		Solid Tissue Normal		
	Female (n=11)	Male (n=7)	Female (n=11)	Male (n=7)	P-value
Richness					
Mean (SD)	20.5(8.1)	21.1(8.8)	32.7(13.5)	23.7(14.9)	
Median [Min, Max]	20.0[9.0,35.0]	26.0[10.0, 31.0]	33.0[18.0,54.0]	17.0[9.0,44.0]	0.008
Diversity (SW)					
Mean (SD)	1.19(0.923)	1.49(0.908)	1.41(0.885)	1.29(0.935)	
Median [Min, Max]	0.75 [0.11,2.6]	1.50 [0.15, 2.57]	1.24 [0.20,2.65]	1.23 [0.11,2.30]	0.623
Evenness					
Mean (SD)	0.387 (0.278)	0.493(0.276)	0.413(0.259)	0.409(0.278)	
Median [Min, Max]	0.25 [0.040, 0.770]	0.60 [0.06, 0.79]	0.35 [0.07, 0.77]	0.44 [0.05, 0.74]	0.706

Table displays richness (number of species) diversity (Shannon-Wiener index of diversity) and Shannon evenness stratified by sample type and sex.

3.1.4.6.4 Cancers of the lung

Sequencing files from tumor and adjacent normal specimen pairs corresponding to 421 cases of lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) were analyzed. Of these, 23 cases in LUSC and 17 cases in LUAD had one or both pairs without microbial reads or only viral like reads. Clinical data was not available for 26% of LUAD cases (52 /200 cases). Demographic characteristics of both patient populations are summarized in **Table 9**. Overall 52% were male and 12% did not report or had missing data. Mean age at diagnosis was 66.8 years (±9.3 SD) with majority self-reporting as being Non-Hispanic White in both cancer cohorts (67% and 81% in LUSC and LUAD respectively). Stratified by sex and sample type, there were no significant differences in age at diagnosis, race, ethnicity, primary diagnosis, tumor stage, or survival days. In LUSC 43% were classified as stage II-III compared to 30% in LUAD. All available samples with bacterial like reads were examined for bacterial composition. Paired tumor and adjacent normal samples were used for correlation analyses.

In LUSC, a total of 9,622 reads in primary tumor and 50,630 reads in adjacent normal were mapped and aligned to 1049 unique OTUs. From these 51% of the tumor reads were identified as viral like reads compared to 15% in adjacent normal. Mean bacterial reads per sample was 107 with approximately 7 bacterial species found per sample. We identified 12 phyla, 26 class, 65 order, 132 family, 310 genus, and 741 unique species. From the 741 species, 335 were shared in tumor and adjacent normal with 80 unique species found in the tumor tissue and 326 in the adjacent normal respectively (**Table 10**). In analyses of variance total number of unique species was significantly lower in tumor compared to the number of shared species and unique species in adjacent normal (F=655.7, p=0.0001). Based on the significant number of reads mapped/aligned to viral reads we wanted to explore the differences in distribution across tumor and adjacent normal. We found a total of 46 unique species corresponding to a number of Torque-teno-viruses, HHV-4 and other Herpesviridae species, where 26 of these were shared between tumor and adjacent normal tissue. There was a non-significant difference in the total number of viral reads in tumor compared to normal and no difference in the number of unique or shared viral species

between tumor and adjacent normal. Several bacterial species reads were found consistently throughout samples belonging to Proteobacteria, Firmicutes and Actinobacteria phyla. The most abundant taxa in the population was Bacillus subtilis, present in 55% of tumors and 56% of the adjacent normal, yet the highest number of reads were identified as *Sphingomonas* spp., present in <20% of the sample population. Similarly, Proteobacteria, Actinobacteria and Firmicutes were the most abundant phyla in LUAD.

The proportion of viral like reads in LUAD was smaller compared to LUSC cohort, contrary to LUSC it was lower in tumor compared to adjacent normal 5 fold. We identified 820 unique bacterial species classified among 12 phyla, 27 class, 35 order, 131 family, and 306 genus. From the 820 species, 449 were shared in tumor and adjacent normal with 190 unique species found in the tumor tissue and 181 in the adjacent normal respectively. **Table 10**. There was no difference in the number of unique species or shared species between tumor and adjacent normal. Like in LUSC, the most abundant taxa in the LUAD sample population was *Bacillus subtilis*, present in 53% of tumors and 63% of the adjacent normal and a total of 110 reads across. Highest number of reads in LUAD were among *Cutibacterium acnes* (1282 reads present in 25% of tumors and 22% of adjacent normal), *Mitsuria* sp7 (1223 reads present in 26% tumors and 29% of adjacent normal), and *Delftia acidovorans* (1093 reads present in 14% of tumors and 17% of adjacent normal). (Supplemental **Table A.1** Core taxonomy across cancer types).

Figure 16 *LUSC Log 2 fold change of significant taxa*

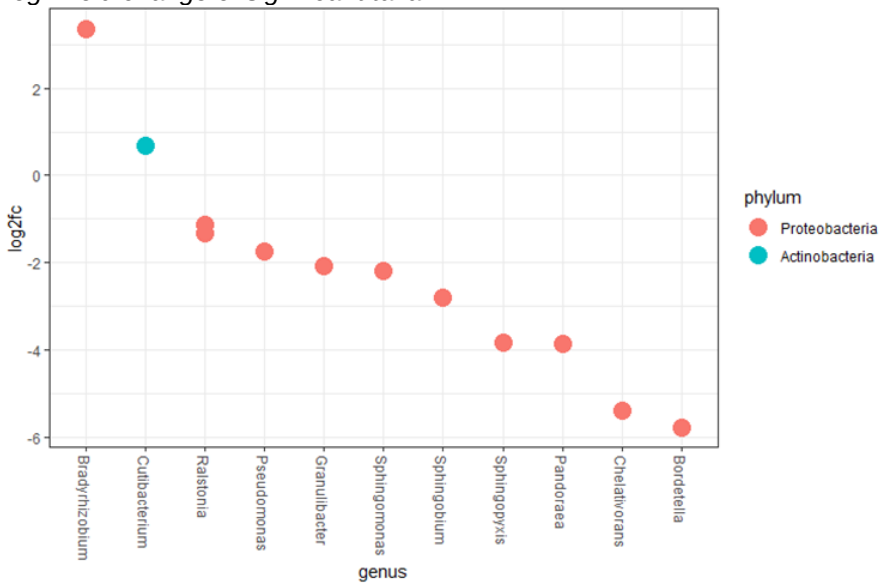


Figure shows fold change (Log2) differential abundant taxa collated at species name in LUSC. Each dot represent a species within the genus, colored by phylum. Species above zero are higher in tumor, below are higher in adjacent normal. Proteobacteria were the most predominant phyla with 11 species. No significant taxa was identified after FDR multiple test correction.

We compared the species composition within tumor and within adjacent normal in LUSC and LUAD. In Wilcoxon rank sum test 54 species were identified in LUSC tumor and 64 in its adjacent normal to have ranks significantly different (FDR<0.05). In LUAD 89 species were identified to have ranks significantly different across samples within tumor (FDR<0.05) and 79 species in adjacent normal. In paired Wilcoxon signed rank test, comparing differential relative abundance between tumor and adjacent normal we found 12 taxa in LUSC (**Figure 16**) and 22 taxa in LUAD (**Figure 17**) to be differentially abundant between tumor and their paired adjacent normal with pvalue <0.05, however there was no difference after correcting for multiple testing (FDR=1).

Figure 17 *Log Fold Change (l2fc) of top taxa in LUAD cohort*

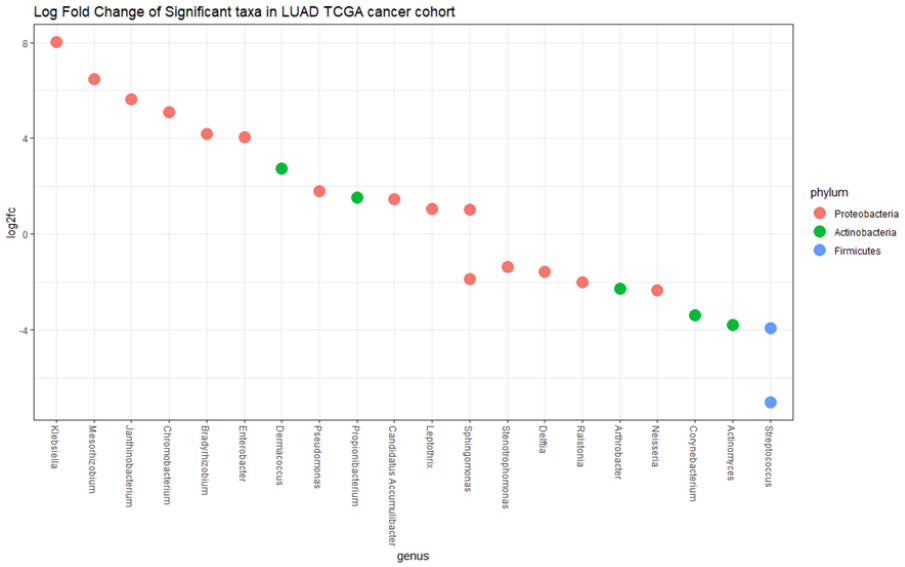


Figure shows fold change (Log2) differential abundant taxa collated at species name in LUAD with pvalue <0.05. Each dot represent a species within the genus, colored by phylum. Species above zero are higher in tumor, below are higher in adjacent normal. Proteobacteria were the most predominant phyla with 15 species. No significant differential abundance was detected (FDR pvalue adjustment > 0.05).

We observed a shift in the ratio of Proteobacteria to Actinobacteria phyla composition in LUSC (log2 P|A tumor =2.63, log2 P|A adjacent normal=5.13) while compositional shift changes in LUAD were negligible (log2 P|A tumor =2.30 and log2 P|A adjacent normal=2.40). **Figure 9**. We then asked if there were differences within sample diversity within and between tumor and adjacent normal samples within each lung cancer cohort. When stratifying by sex and sample type, there were no statistically significant differences in species richness, within sample alpha diversity (measured by Shannon-Wiener Index), or the evenness spread (defined as alpha diversity / log normal (species richness)) in either cohort overall. There were observable differences in alpha diversity by age at diagnosis groups, tumor stage. **Table 14**. Younger female patients at diagnosis in the LUSC cohort, had higher within sample diversity in tumor

compared to adjacent normal. While LUAD, when looking at tumor staging, females had higher diversity in adjacent normal compared to tumor although these were not significant in analyses of variance.

Table 14 *Richness and Diversity in lung squamous cell carcinoma and lung adenocarcinoma*

LUSC	**Primary Tumor N=221**		**Solid Tissue Normal N=221**		
	Female N=64	**Male N=157**	**Female N=64**	**Male N=157**	**Pvalue**
Age at diagnosis Mean (SD)	68.2 (8.19)	68.4 (8.51)	68.2 (8.19)	68.4 (8.51)	
Diversity Overall Mean (SD)	1.83 (0.845)	1.64 (0.943)	1.85 (0.934)	1.76 (0.990)	
Richness Overall Mean (SD)	11.1 (17.6)	9.34 (11.3)	12.4 (15.9)	14.8 (37.8)	
Evenness Overall Mean (SD)	0.932 (0.145)	0.877 (0.250)	0.891 (0.225)	0.869 (0.245)	
Age at Diagnosis group					
40-50 (n=6)	2.29 (0.34)	2.13 (0.41)	1.30 (0.74)	2.32 (0.77)	*
51-60 (n=31)	1.97 (0.70)	1.48 (1.09)	1.67 (0.52)	1.93 (0.80)	
61-70 (n=83)	1.44 (0.82)	1.59 (0.82)	1.75 (0.79)	1.64 (0.98)	
71-80 (n=85)	2.15 (0.77)	1.70 (0.97)	2.18 (0.99)	1.81 (1.06)	
>80 (n=16)	2.13 (0.69)	1.66 (1.07)	1.33 (1.46)	1.54 (0.83)	0.078
Tumor Stage					
Stage I (n=121)	1.78 (0.89)	1.61 (0.95)	1.77 (0.92)	1.72 (0.96)	
Stage II (n=56)	1.84 (0.68)	1.58 (0.98)	2.14 (0.73)	1.84 (0.95)	
Stage III (n=40)	2.00 (0.79)	1.83 (0.84)	1.82 (1.04)	1.72 (1.08)	
Stage IV (n=3)	--	1.23 (0.39)	--	1.49 (1.05)	
Not Reported (n=1)	--	2.34 (0.00)	--	3.18 (0.00)	0.342

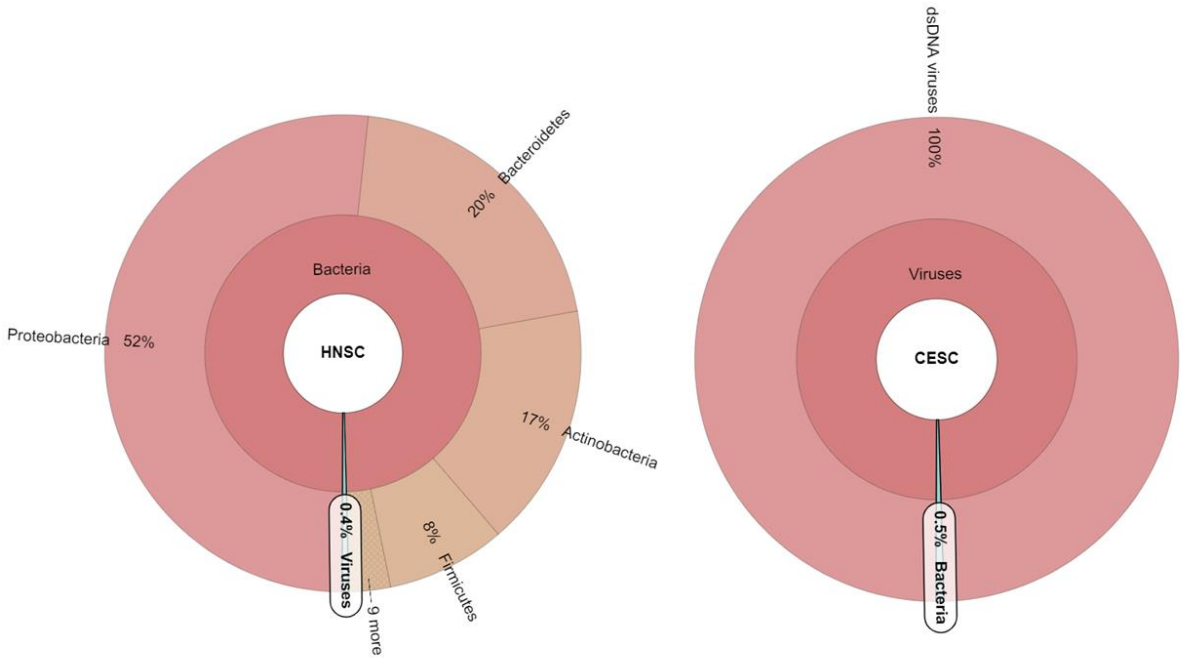
LUAD	**Primary Tumor N=137**		**Solid Tissue Normal N=137**		
	Female N=77	**Male N=60**	**Female N=77**	**Male N=60**	**Pvalue**
§Age at diagnosis Mean (SD)	64.7 (10.5)	64.4 (9.49)	64.8 (10.5)	65.0 (9.09)	
Diversity Overall Mean (SD)	1.8 (1.06)	1.9 (0.85)	1.8 (0.75)	2.0 (0.72)	
Richness Overall Mean (SD)	13.7 (18.4)	10.5 (14.3)	11.4 (11.6)	10.7 (7.5)	
Evenness Overall Mean (SD)	0.936 (0.10)	0.969 (0.04)	0.925 (0.12)	0.950 (0.09)	
Age at Diagnosis group					
40-50 (n=21)	1.91 (0.90)	2.19 (0.70)	1.57 (0.55)	1.63 (1.18)	
51-60 (n=58)	1.84 (1.10)	1.64 (1.11)	2.04 (0.88)	1.84 (0.79)	
61-70 (n=94)	1.90 (1.14)	1.88 (0.94)	1.73 (0.81)	1.92 (0.68)	
71-80 (n=71)	1.53 (0.94)	1.80 (0.51)	1.99 (0.46)	2.03 (0.59)	
>80 (n=12)	1.88 (0.67)	1.99 (0.09)	1.38 (0.43)	2.56 (0.16)	0.482
Missing	8 (11%)	10 (18%)	8 (11%)	10 (18%)	
Tumor Stage					
Stage I (n=143)	1.67 (1.10)	1.93 (0.85)	1.84 (0.77)	1.95 (0.72)	
Stage II (n=66)	2.29 (0.98)	1.77 (0.77)	1.60 (0.80)	2.01 (0.83)	
Stage III (n=50)	1.85 (0.84)	1.74 (0.99)	2.09 (0.61)	1.91 (0.59)	
Stage IV (n=11)	1.55 (1.09)	2.17 (0.48)	2.12 (0.23)	1.94 (0.55)	
Not Reported (n=4)	1.98 (0.00)	1.74 (0.00)	2.04 (0.00)	2.71 (0.00)	0.933

LUSC (top) and LUAD (bottom) diversity measures in patients of different age groups at diagnosis and different tumor staging stratified by sample type and sex. Table shows global pvalue with significance codes (0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1) for individual variables. -- data not available. In LUSC the 40-50 age group has significant differences within the female group comparing tumor to adjacent normal and when comparing within tissue type females to males adjacent normal.

3.1.4.6.5 HPV associated cancers of the head & neck and cervical cancer

We analyzed 69 squamous cell carcinoma of the head and neck region (HNSC) and 8 cervical squamous cell carcinoma specimens from TCGA. Demographic characteristics of patient population with available clinical data are summarized in **Table 9**. There were significantly more males than females (70% vs 30%) in our subset of paired HNSC sample population. Because the established association with HPV etiology, we examined viral read presence in both cancers. Overall we detected viral like reads in 14% of HNSC (20/143) and 100% of CESC samples (16/16), while bacterial like reads were detected in all samples for both cancer types (**Table 8**). We found that proportion of bacteria versus viral reads for HNSC and CESC cancers were opposite to each other. In HNSC bacterial reads accounted for more than 99% of the total reads while in CESC bacterial like reads were less than 1% of the total (**Figure 18Figure 1**).

Figure 18 *Microbial composition in HNSC and CESC*

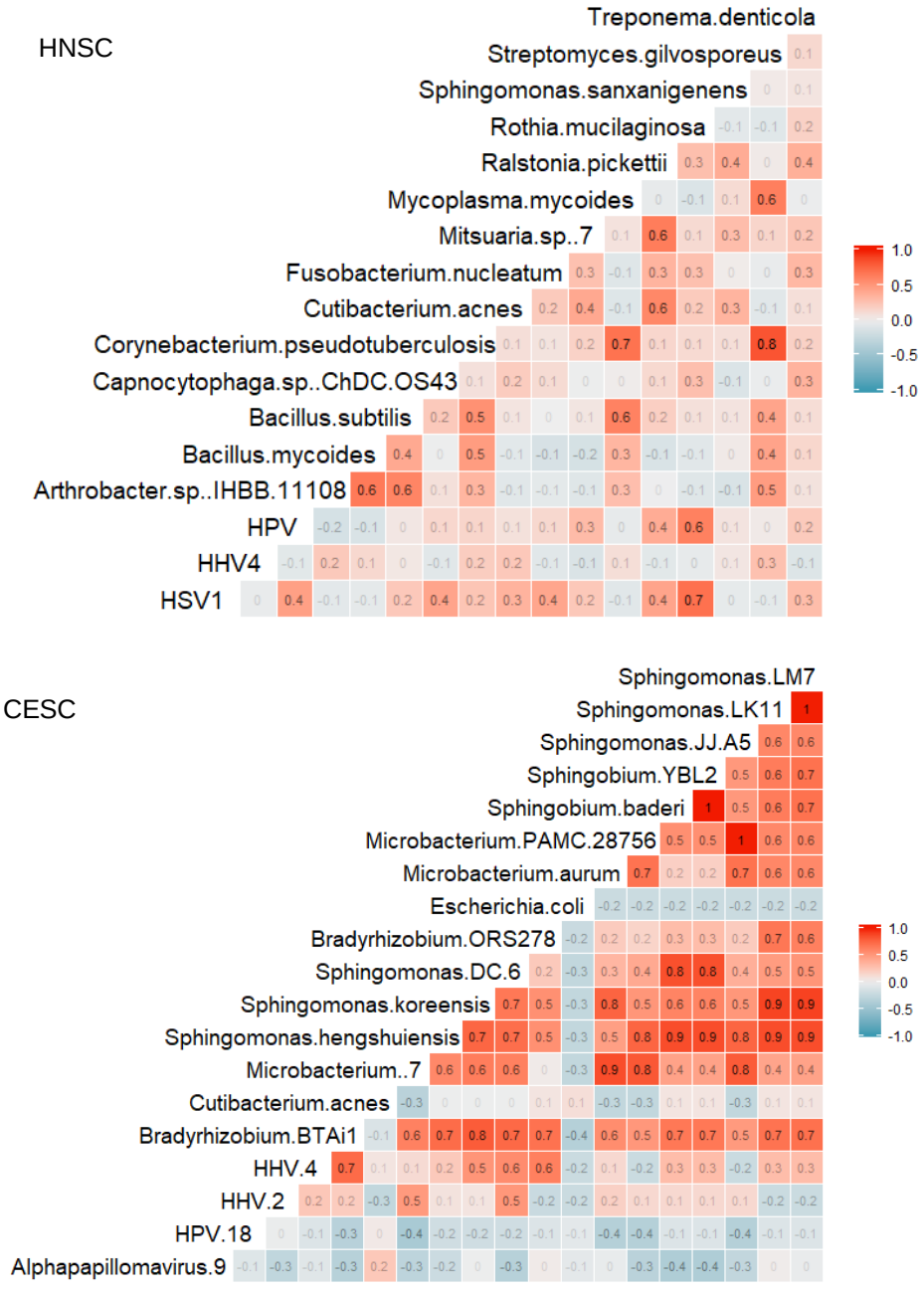


Krona plots of HNSC and CESC cancer cohorts total reads. HNSC and CESC had inversed proportion of viral and bacterial presence. Krona plots are colored clock-wise (6th hour mark) from red to green tones in descending order of total reads detected, slices show percent proportion of major phyla.

HPV like reads were detected within the HNSC cohort in agreement with previous findings (Tang et al. 2013, Wang et al. 2017). HHV-4 was detected in approximately 13% of HNSC samples. HNSC specimens with positive detection of HPV like reads were not found to be co-infected with HHV-4. Contrary to that in CESC HPV and HHV-4 were found to co-occur in several samples. Based on casual, epidemiological and meta-analysis data (Zhu et al. 2016) linking *Chlamydia trachomatis* co-infection to susceptibility to cervical cancer after HPV infection, we evaluated co-occurrence of the bacterium and

HPV in CESC. We found no evidence of *Chlamydia trachomatis* reads in our subset of CESC samples, interestingly we found evidence in HNSC tumor samples. The presence of HPV was correlated with presence of top abundant taxa in HNSC and CESC (**Figure 19**).

Figure 19 *Microbial presence correlation matrix in HNSC and CESC cohorts*



Correlation matrix. Matrices show Spearman correlation coefficients. Color depth indicate strength and significance of the correlation. Color gradient from red, positively correlated to blue, negatively correlated. In HNSC, HPV is correlated with presence of alpha herpesvirus-1 and with *R. mucilaginosa.* In CESC HPV has a negatively trend with bacterial presence

No significant correlation was found between HPV status and diversity or bacterial abundance in CESC. HPV status was correlated with microbial abundance HNSC.

We examine differential abundance within tissue type and between tumor and adjacent normal paired sample within each cohort. Relative abundance was not significantly different for either tumor type compared to adjacent normal. In HNSC, there were no differences in bacterial diversity means in paired tests comparing tumor to adjacent normal (p=0.6). Slight differences were observed by anatomical site among HNSC larynx, LOP (lip, oral and pharynx overlap), and tongue (base and non-specified) when compared against floor of mouth within each tissue type when stratifying by sex and sample type. In analyses of variance, anatomical site was a predictor of alpha diversity ($p < 0.001$). When stratifying by sample type and sex, alpha diversity was slightly higher among females adjacent normal sample with significant differences by anatomical site (p=0.002) after controlling for sample type, tumor stage, smoke, age, sex and race (**Table 15**).

Table 15 *Diversity measures in HNSC anatomical sites and tumor stage by sample type and sex*

HNSC	**Primary Tumor** N=69		**Solid Tissue Normal** N=69		
	Female n=21	Male n=48	Female n=21	Male n=48	**Pvalue**
Mean (SD)	1.66 (0.79)	1.80 (0.90)	1.99 (0.91)	1.95 (0.86)	
Median [Range]	1.70 [0, 3.22]	1.82 [0, 4.08]	2.21 [0, 4.04]	1.99 [0, 3.84]	
Anatomical Site					
Floor of mouth (n=6)	2.27 (0.00)	3.24 (0.85)	2.46 (0.00)	2.88 (0.43)	
Hard palate (n=2)	3.22 (0.00)	--	4.04 (0.00)	--	
Larynx (n=46)	1.31 (0.42)	1.61 (1.06)	1.53 (0.25)	1.95 (1.08)	
Lip, oral & pharynx (n=34)	1.79 (0.63)	2.0 (0.44)	2.05 (0.62)	1.69 (0.70)	
Tongue, Base (n=6)	--	1.69 (0.28)	--	1.92 (0.50)	
Tongue, NOS (n=44)	1.33 (0.79)	1.77 (0.67)	1.74 (1.04)	1.99 (0.50)	**0.002**
Tumor Stage					
Stage I (n=2)	2.62 (0.00)	--	2.21 (0.00)	--	
Stage II (n=36)	1.27 (0.56)	1.84 (0.59)	1.61 (0.46)	1.62 (0.83)	
Stage III (n=24)	1.43 (0.85)	2.11 (0.92)	2.42 (0.07)	2.36 (0.54)	
Stage IV (n=76)	1.89 (0.72)	1.70 (0.96)	2.02 (1.17)	1.98 (0.88)	0.461

Table shows alpha diversity index (Shannon-Weiner index of diversity) across different anatomical sites and tumor stages in HNSC cohort stratified by sample type and sex. Significant figures (pvalue) are bolded. –NA

Measures of evenness and species richness (diversity index) were measured by various methods and good correlation by pairwise analysis between these methods was observed when evaluating each tissue site separately and in paired analyses. In general diversity index means in tumor were lower compared to adjacent normal tissue in HNSC. There was no difference in CESC diversity index means between tumor and adjacent normal samples (p=0.11). We note, that samples with lower diversity index in tumor compared to adjacent normal were more likely to be of the atypical pleomorphic squamous cell subtype (log odds= 2.19). There were no differences by tumor stage.

3.1.4.6.6 Bladder

We examined 850 files from 412 BLCA cases. From these, most were technical replicates. A total of 56 paired tumor and adjacent normal samples sequences were selected for further evaluation at a 1:1 ratio. The majority of the cases were male (68%), White (89%) and 92% non-Hispanic. Mean age at diagnosis was 69 years (±10.7 SD) and 50% were classified at pathological tumor stage IV (**Table 9**). All cases were positive for at least 1 bacterial read at any detection level. Within these cases, we identified 7 phyla, 12 classes, 31, order, 60, family, 105 genus and 195 unique bacterial species (**Table 10**). Overall, Proteobacteria were the most abundant species, making 93% of the total reads with *Stenotrophomonas maltophilia* the most abundant species (61% of the total reads). However, prevalence within the sample population was low. Positive detection was identified in 3 tumor and 1 adjacent normal.

Figure 20 *BLCA overall microbial composition*

Krona plot of microbial composition among 7 taxonomic levels in BLCA tumor and adjacent normal. Plot shows proportion of bacterial and viral reads present in all samples combined magnified to highest taxonomy level. In BLCA 93% of the reads were identified as Proteobacteria, with *Stenotrophomonas maltophilia* reads making 61% of these. *Stenotrophomonas maltophilia* was detected in 3 tumor and 1 normal samples. This highlights measurements of prevalence must be taken into account for accurate interpretation.

There was a significant shift in the total number of reads classification at the phylum level in tumor compared to adjacent normal (**Figure 3.2B**). BLCA adjacent normal reads were distributed among Proteobacteria (50%), Firmicutes (25%), Actinobacteria (20%), while almost all (98%) of the tumor reads were from the Proteobacteria phyla. Despite, when considering total number of reads, relative abundance and percent positivity (presence above relative abundance of 0.2% threshold), we found no statistically

significant differences between paired tumor and adjacent normal samples. We compared the relative abundances by Wilcoxon rank sum test within tumor and within adjacent normal and signed rank test between sample types. No taxa were found to be differentially abundant in paired analyses after multiple test correction. However, we note that *Cutibacterium acnes* reads were uniquely identified in BLCA tumor samples (p=0.03 FDR=1, L2FC= 3.1). Because the large number of non-paired files filtered out, we completed unpaired differential analyses (PathoStat-edgeR function). When considering all available tumor and adjacent normal files regardless of 1:1 pairing, several species were found to be differentially abundant in tumor compared to normal after FDR multiple test correction including *Mathylobacterium radiotolerans, Pseudomonas aeruginosa* and *Pseudomonas putida* identified to be higher in adjacent normal (p= < 0.001, FDR = < 0.05, LFC=1.61, 2.99 and 1.65 respectively) and *Cupriavidus metallidurans* higher in tumor (p= < 0.001, FDR = < 0.05, logFC -3.47). *Stenotrophomonas maltophilia* was not identified as differentially abundant in paired or unpaired analyses. With the exception of two cases originating from

Table 16 *Diversity measures in BLCA anatomical sites, age at diagnosis and tumor stage stratified by sample type and sex*

	Primary Tumor N=28		**Solid Tissue Normal** N=28		
	Female n=9	Male n=19	Female n=9	Male n=19	**Pvalue**
Overall Diversity (SW)					
Mean (SD)	1.21 (0.452)	1.30 (0.611)	1.16 (1.09)	1.37 (0.647)	0.895
Median [Range]	1.17 [0.693, 1.75]	1.13 [0.00, 2.31]	0.693 [0.00,2.51]	1.39 [0.00,2.57]	
Anatomical Site	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	
Anterior wall (n=4)	--	1.31 (0.19)	--	1.25 (0.36)	
Dome (n=6)	1.61 (0.00)	1.68 (0.64)	2.04 (0.00)	1.77 (0.39)	
Lateral wall(n=6)	1.75 (0.00)	1.55 (0.49)	0.00 (0.00)	1.11 (0.42)	
Posterior wall (n=6)	--	0.97 (0.29)	--	1.16 (0.18)	
Trigone (n=10)	1.16 (0.44)	1.42 (0.32)	0.46 (0.33)	1.63 (0.94)	
Bladder, NOS (n=24)	1.01 (0.33)	1.24 (0.71)	1.76 (1.02)	1.37 (0.71)	0.495
Age at Diagnosis group					
40-50 (n=2)	--	2.21 (0.00)	--	1.79 (0.00)	
51-60 (n=10)	1.46 (0.24)	1.68 (0.64)	1.54 (1.10)	1.77 (0.39)	
61-70 (n=16)	--	1.10 (0.62)	--	1.23 (0.63)	
71-80 (n=18)	1.11 (0.38)	1.32 (0.42)	0.91 (0.85)	1.22 (0.67)	
>80 (n=10)	1.04 (0.50)	1.24 (0.41)	1.03 (1.01)	1.74 (0.36)	0.277
Tumor Stage					
Stage I (n=0)	--	--	--	--	
Stage II (n=6)	--	1.33 (0.52)	--	1.25 (0.61)	
Stage III (n=22)	1.39 (0.30)	1.48 (0.54)	1.48 (0.81)	1.30 (0.78)	
Stage IV (n=28)	1.06 (0.46)	1.16 (0.62)	0.90 (1.11)	1.46 (0.47)	0.567

BLCA diversity measures in samples from diverse anatomical sites, patients of different age groups at diagnosis and different tumor staging stratified by sample type and sex. SW: Shannon-Wiener index of diversity. –NA

same institution, these taxa did not often co-occurred. Coincidentally, cases identified with co-occurrence were also identified to drive the proportion of *Stenotrophomonas maltophilia* presence in BLCA cohort.

Presence of a combination of these species could be indicative of disease status or sepsis. Bacterial alpha diversity was associated with age at diagnosis in BLCA (rho =-0.32, p=0.02). In analyses of variance age at diagnosis was an independent predictor of alpha diversity (F= 4.06, p =0.049) when controlling for sample type, race and sex (**Table 16**). In fully adjusted model, there are no significant differences in alpha diversity. Likewise although alpha diversity means were higher in adjacent normal overall, no statistical significant differences were noted in paired tests by sex, race, anatomical site or tumor stage.

3.1.5 **Validation of bacterial species in gastric and lung cancers**

To validate our methods and bioinformatics detection of significant species we selected stomach and lung cohorts based on known infectious and no known infectious etiological factors. Bioinformatics findings were validated with tissue from an independent population by species-specific qPCR. Institutional Review Board approval was obtained from the University of Hawaii IRB. Paired tumor and adjacent normal FFPE samples from the Hawaii Tumor Registry-Discard Residual Repository (Hawaii RTR) were requested at a similar 1:1 ratio per case. DNA was extracted using appropriate purification kit to maximize yield (Methods). We selected 21 gastric cases and 60 lung adenocarcinoma cases on the basis of paired tumor and adjacent normal availability. Quantitative PCR reactions were performed per methods section in duplicate. Duplicate discrepancies were retested.

In TCGA STAD cohort, *Selenomonas sputigena* had the highest proportion of mapped reads detected in tumor samples compared other species with a prevalence of 18% of tumor compared to 9% of adjacent normal samples. Its presence was correlated with presence of *Fusobacterium nucleatum* in tumor (rho=0.48, p=<0.001) and with *Helicobacter pylori* in adjacent normal (rho=0.26, 0.01). There was a correlation between *Selenomonas sputigena* presence and HHV-4 status (tumor: rho=0.24, p=0.05 adjacent normal: rho=0.31, p=0.004). In the other hand, *Helicobacter pylori* was found to be differentially abundant between tumor and adjacent normal pairs with significant higher prevalence in adjacent tissue compared to tumor samples (n=23 and n=7, respectively), whereas *Fusobacterium nucleatum* was detected at very low abundance levels uniquely identified in 8% of tumor samples. We therefore wanted to validate detection of *Selenomonas sputigena*, *Fusobacterium nucleatum* and *Helicobacter pylori* in in tumor and adjacent normal specimens. Of the Hawaii RTR gastric cases, 76% were over the age of 60 (n=16), 62% were females (n=13), 95% were other than White (n=20), 62% were classified as poorly differentiated grade III (n=13), 24% were from the gastric antrum (n=5) and 43% non-specified (n=9). Compared to our subset of TCGA gastric cancers (n=85) where 75% were over the age of 60 (n=62), 44% were females (n=37), 36% were other than White (n=31), 56% were classified as grade III (n=48), and 31% were excised from the gastric antrum (n=27).

Overall, positive detection in RTR dataset was almost half that of TCGA-STAD dataset (27% vs 47%). Compared to STAD, in RTR gastric samples there was no association between bacterial presence and tissue type. We believe this could be due to the small sample size in Hawaii RTR compared to TCGA-STAD and the racial composition difference were TCGA-STAD were mostly White compared to Hawaii RTR, mostly Asian ethnic subgroups and Hawaiian. One of the most significant findings is the, detection of species presence by qPCR in with similar patterns of co-occurrence. *Selenomonas sputigena* was detected in 10% (n=2) of tumor samples, *Fusobacterium nucleatum* and *Helicobacter pylori* in were detected in 14% (n=3) each. Two cases were positive for *Selenomonas sputigena* and *Fusobacterium nucleatum*. In RTR dataset, bacteria presence in tumor of *Helicobacter pylori*, *Fusobacterium nucleatum* and *Selenomonas sputigena* was associated with tumor stage and anatomical site.

Table 17 *Proportions tables predicted versus observed in gastric cancer*

Helicobacter pylori N			
T	(+)	(-)	
(+)	4 (17)	3 (5)	7(8)
(-)	19 (82)	59 (95)	78 (92)
	23(27)	62 (73)	85 (100)
Chi-sq	10.2		
p	0.001		

Selenomonas sputigena N			
T	(+)	(-)	
(+)	3 (3)	12 (19)	15 (18)
(-)	5 (6)	65 (84)	70 (82)
	8 (9)	77 (91)	85 (100)
Chi-sq	2.1		
p	0.146		

Fusobacterium nucleatum N			
T	(+)	(-)	
(+)	0 (0)	8 (9)	8 (9)
(-)	0 (0)	77(91)	77(91)
	0 (0)	85 (100)	85 (100)
Chi-sq	6.1		
p	0.001		

Helicobacter pylori N			
T	(+)	(-)	
(+)	1 (50)	1 (5)	2 (10)
(-)	1 (50)	18 (95)	19 (90)
	2 (10)	19 (90)	21 (100)
Chi-sq	0		
p	1		

Selenomonas sputigena N			
T	(+)	(-)	
(+)	0 (0)	2 (10)	2 (10)
(-)	0 (0)	19 (90)	19 (90)
	0 (0)	21 (100)	21 (100)
Chi-sq	0.5		
p	0.48		

Fusobacterium nucleatum N			
T	(+)	(-)	
(+)	2 (10)	1 (5)	3 (14)
(-)	0 (0)	18 (86)	18 (86)
	2 (10)	19 (90)	21 (100)
Chi-sq	0		
p	1		

Association between bacterial presence in tumor and adjacent normal in STAD (predicted) and RTR (observed) detection proportions. Table shows count of positive samples and proportions (%) for *Helicobacter pylori*, *Selenomonas sputigena* and *Fusobacterium nucleatum* with McNemar's (chi-square) test and pvalues. There was an association between tissue type and bacterial presence for *Helicobacter pylori* and *Fusobacterium nucleatum* in STAD while no association was found in RTR gastric samples.

Compared to STAD, *Selenomonas species* had very low abundance and were uniquely detected in lung adenocarcinoma adjacent normal tissue (upper lobe) of patients classified at stage I. Presence of *Selenomonas* species was similarly validated. We selected tumor and adjacent normal samples from 60 lung adenocarcinoma cases from the Hawaii RTR. These cases 75% (n=45) were over the age of 60), 27% (n=16) White), 33% (n=20) classified at stage I. For convenience, we utilized commercially available *Selenomonas* primers (*Selenomonas sputigena*). Similar to TCGA-LUAD cohort, in Hawaii RTR dataset, *Selenomonas* was uniquely identified in adjacent normal tissue. Our results confirmed *Selenomonas* was uniquely in adjacent tissue. Contrary to TCGA bioinformatics detection in the upper lobe, Hawaiii RTR population detection was in the lower lobe. Perhaps this is related to inter-individual differences and further studies are needed to confirm. These results suggest that bacteria presence as well as bacterial composition differences in tumor tissue can be predicted from whole exome sequencing data to

determine clinically relevant species. Continued detection of other species at different abundance levels with larger population size should further confirm our findings.

3.1.6 **Discussion**

This study showed differences in microbial composition in paired tumor and adjacent normal tissue sequencing samples across 9 TCGA cancer cohorts. Through our microbial detection methods we showed differential bacterial abundance in stomach and colon adenocarcinomas. Our findings are consistent with other reports. Further, we used viral detection as internal validation and our findings which are consistent with previous reports. The role of *Helicobacter pylori* in cancer has been firmly established in stomach adenocarcinoma we add potential interaction with microbial community in the adjacent tissue as sign of disease progression. We note that measures of relative abundance alone or total number of reads do not provide sufficient information regarding the compositional differences in the tumor microenvironment. Measures of total reads, relative abundances and prevalence in the population need to be taken into account for a more accurate description of the differences within and across cohorts. In LIHC we determined that there was no difference in bacteria composition within paired samples when comparing tumor to its adjacent normal tissue, however there was an observable difference within tissue type in the cohort. Moreover although the number of species per sample appeared to be similar both tumor and adjacent normal, diversity varied by stage, age at diagnosis and sex which could have potential clinical significance particularly when we seek to uncover targetable biomarkers to improve patient outcomes. We identified HPV reads in all CESC paired samples as previously reported (Tang et al. 2013, Cantalupo, Katz, and Pipas 2018). In cancers of the head and neck, HPV was detected in 5% of the samples. Although previously reported detection of HPV for this group of cancers ranges from 20% to 21% (Cantalupo, Katz, and Pipas 2018, Khoury et al. 2013, Hernandez et al. 2014) we feel confident that our results are similar to those previously reported using WXS data. Our study included a very small sample of 69 paired cases. From these 7 samples (6 tumor, 1 adjacent solid tissue normal) corresponding to 6 cases (1 female, 5 male) were positive for HPV. In CESC we found a (weak) negative correlation between HPV and *Bradyrhizobium* sp. which varied by pathological stage. We note that our sample size was small; perhaps correlation may be more prominent with increased sample size. Interestingly, Riley et al. reported that Bradyrhizobium like species including Bradyrhizobium BTAi1 were the most-common strain level operational taxonomic units found within the 1000 Genomes Project which supported lateral gene transfer (Riley, 2013). Laurence et al. pointed out that Bradyrhizobium sp. were found to be common contaminants due to ultrapure water systems within the 1000 Genome Project and many high-throughput efforts (Laurence et al. 2014 PMID 24837716). Riley, notes that although contamination can be suspected, presence of the microbe may be due to diet and lifestyle differences in the population, highlighting that little is known about the composition of the human microbiome. Riley et al. utilized data derived from Chinese population. Microbial profiles in this study were derived from TCGA

project which encompasses different racial groups, primarily Caucasian Americans. Data has been collected at various Institutes, and was sequenced at different Centers. Although, water system or laboratory contaminants could be a source of reads, strict use of paired samples should assist with misidentification. We note that presence varied among cohorts with highest total reads among CESC (originating from University of Washington) and COAD (originating from Christiana Healthcare and Indivumed). CESC numbers of reads in tumor were 22 times higher compared to adjacent normal, while number of *Bradyrhizobium* like reads in COAD tumor were 7 times that of adjacent normal. We also note that often both case pairs had bacterial reads for *Bradyrhizobium* like species. *Bradyrhizobium diazoefficiens* and *Bradyrhizobium* BTAi1 are nitrogen fixating bacteria and their role in disease is unknown. Nevertheless, these species may have a potential role in cancer pathogenesis and cancer therapy by means of their Hsp70 family molecular chaperone protein interaction with p53 (Deocaris et al. 2007, Shevtsov, Huile, and Multhoff 2018). Our study is not without its limitations, the low reads relative to human sequences may not be sensitive to magnitude of differential expression and it may be less powered because our paired analyses filtration which resulted in a low number of cases analyzed. We sat limits to protect against this by not including any cancer cohorts with less than 15 specimens (smallest sample size CESC with 16 specimens). Our integrated analysis of exclusive 1-to-1 paired samples is not sensitive to tissue specific baseline relative abundance, or inherited 16S compositional assumptions. Each patient served as their own control eliminating interacting and confounding factors. In addition we provide simple and easy to interpreted results for complicated microbial data across a subset of TCGA cancer cohorts. We conclude that identifying microbial composition in tumor and adjacent normal tissue, using whole exome sequencing data provides useful and comparative tool similar to transcriptome and metagenomic methods to study bacterial composition in cancer. Further qPCR validation of bacterial presence with tissue specimens from Hawaii Tumor Registry as an independent population strengthens our findings. We highlight co-occurrence of *Selenomonas sputigena* and *Fusobacterium nucleatum* in tumor tissue of stomach adenocarcinoma. These oral species have been identified in the tongue coating of gastric patients but have not been identified in the tumor tissue previously (Xu et al. 2019). Future studies seeking to characterize the microbiota within the tumor microenvironment should consider examination of the adjacent tissue weighing prevalence within the population with equal weight to the total amount of reads detected. This will facilitate microbial functional predictions and distinguish between true presence and laboratory artifacts and possible contamination.

3.1.7 **Materials and Methods**

3.1.6.2 Cancer Database

The data used in this study were derived from the TCGA consortium (phs000178 versions v9.p8 and v10.p8 under Project-14778, Deng, PI). TCGA cancer types with whole exome sequencing case pairs meeting selection criteria were downloaded. Cases were defined as solid tumor cancer types within TCGA that had human aligned sequencing reads, paired primary tumor and solid tissue normal raw

exome sequences in BAM file format, plus available clinical data. Paired cases were selected at 1:1 ratio for the bioinformatics interrogation

3.1.6.3 Computational Framework for Microbial Detection

TCGA Level-1 data were used to derive microbial information. For microbiota identification we used a bioinformatics pipeline designed to generate microbial profiles from human whole exome sequencing binary version of Sequence Alignment/Map (BAM) files based on PathoScope 2.0. PathoScope 2.0 quantifies microbial strain level proportions found in metagenomics sequencing data (Hong et al. 2014). PathoScope 2.0 pipeline is freely available for download at: http://sourceforge.net/projects/pathoscope/. Quality trim and filtering were completed using SAMtools and Picard. Additional BLAST step was used to subtract any remaining human reads using a custom library. Pipeline produced 3 reports of quantified microbial proportions. Reports were used for bacterial differential analyses. Viral DNA detection o include bacterial phages was used as internal validation. We corroborated viral DNA detection rate, mainly HPV, HBV and HHV-4 (EBV) to previous RNA sequencing works by other authors (). Presence of viral DNA was also used in co-occurrence correlation analyses. R-software was used for phylogenic classification and statistical analyses.

3.1.6.4 Core Microbiota

Identification of core microbiota was completed under study assumptions of relative abundance positivity threshold of ≥ 0.2% per microbe with a minimum prevalence of 20% in the population. Core taxa identification was verified using microbiome R package (version 1.3.3) with default settings and detection and prevalence rates per study assumptions. Any species identified within each cohort were then compared across all cohorts regardless of study assumptions. Visualization of shared taxa was performed with UpSetR package (Alexander Lex et al. UpSet: Visualization of Intersecting Sets, IEEE Transactions on Visualization and Computer Graphics (InfoVis'14), vol.20, no.12, pp.1983-2014. Doi:10.1109/TVCG.2014.2346248).

3.1.6.5 Diversity metrics and Differential abundance analyses

Diversity measurements of alpha diversity (within sample diversity) and beta diversity (between samples) were completed for each cancer type. Mean differences of 15% are considered clinically relevant. Analyses were completed using R-software packages, phyloseq (v.1.25.3) and microbiome (v.1.3.3). Alpha (Shannon-Wiener Index, Simpson Index of Diversity and Fisher's alpha) and Beta diversity were calculated using vegan R package (v.2.5-3). Differential relative abundance were determined using DESeq2 (v.1.21.24), edgeR (v.3.23.5) for count data and Wilcoxon Signed Rank test for compositional data within R-platform. Due low abundance relative read count nature of our data, use of DESeq2 and edgeR packages was limited to few cancer types. Wilcoxon Signed Rank test was used across all cancer types for differential analyses. Bacterial taxa with false discovery rate (FDR) adjusted p-value <0.05 were considered significant at genus and species level. Correlation between the tools is beyond the scope of

this study. We note that all 3 methods identified similar species as differentially abundant, although edgeR on average identified a greater proportion compared to DESeq2 and Wilcoxon Signed Rank test using R. A list of R-packages used is located in Appendix D. **R-Package tools**pp 134.

3.1.6.6 Statistical analyses

To determine the association between differences in relative abundance in tumor and its adjacent normal and clinical features, paired or unpaired t-test and analysis of variance (ANOVA) were used for two- and multi-group comparisons, respectively. Equivalent non-parametric tests were used for non-normally distributed data and to account for compositional structure of microbial relative abundances. Chi-square test was used for categorical data.

3.1.6.7 PCR validation

Institutional Review Board approval from the University of Hawaii was obtained prior to any procedures. We experimentally validated bioinformatics findings with de-identified archival tissue from the Hawaii Tumor Registry-Discard Residual Repository (RTR), a unique collection of formalin-fixed, paraffin-embedded (FFPE) tissue from cancer patients diagnosed within the catchment area of the Hawaii Tumor Registry. The Hawaii Tumor Registry is one of three population-based registries associated with the National Cancer Institute (NCI) and the Surveillance, Epidemiology, and End-Results (SEER) program. Archival tissue from a total of 161 paired cases from gastric (21) and lung (80) cancers were selected for validation. Specimen retrieval, cut & slide, sectioning, pathology review and nucleic acid extraction were performed by the University of Hawaii Cancer Center Pathology Shared Resources. DNA was extracted from FFPE using Qiagen All Prep FFPET Kit (Qiagen, Valencia, CA) and quantified by NanoDrop spectrophotometer (Thermo-Scientific, Wilmington, DE). PCR were completed using 30ng of DNA for every 25µl of reaction mix using commercially available species-specific primer-probe kits (Microbial DNA qPCR assay kits 330033, Qiagen, Valencia CA) per manufacturer's instructions under the following conditions: Activation: 10 minutes 95ºC, followed by 45 cycles of Denaturation and Annealing at 95ºC for 15 seconds and 60ºC for 2 minutes. Samples were tested in duplicates plus positive and negative controls. Discrepancies were resolved by repeat qPCR. Due to budgetary constraints, species-specific validation was limited to *Helicobacter pylori*, *Fusobacterium nucleatum, and Selenomonas sputigena.*

3.1.6.8 Other analyses

KRONA^{TM} plots (https://github.com/marbl/Krona/wiki) were created for relative abundance visualization using Excel macro enable templates. Quantified proportions of bacteria and viruses generated from the bioinformatics pipeline were used to generate plots. Including total per microbe read count, average reads per microbe, percent population prevalence and relative abundance data.

3.1.7 **Acknowledgements**

This work was supported by Ola HAWAII, National Institute on Minority Health and Health Disparities (NIMHD) a component of the National Institute of Health (NIH) grant number 2U54MD007601-32 to

V.S.K. The Bioinformatics Core is supported in part by NIH grant numbers P20GM103466, U54MD007584 and 5P30GM114737. The contents of this work are solely the responsibility of the authors and do not necessarily represent the views of NIMID or NIH.

The results published here are in part based upon data generated by the Cancer Genome Atlas (TCGA) managed by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). Information about TCGA can be found at http://cancergenome.nih.gov. All human data were handled in accordance with TCGA Data Use Certification Agreement and Data Access Request (DAR) 57292; Project-14778 (Y. Deng, PI), Request ID 57292-2 and 57292-4 (renewal 10/10/2018) for access phs000178 versions v9.p8 and v10.p8.

3.1.8 **Literature Cited**

Located in Appendix E, pp.147

3.2 **Bacterial diversity correlates with survival in infection-associated cancers of the head & neck, liver and stomach**

3.2.1 **Abstract**

Objective: One in five cancers are attributed to infectious agents and the extent of the impact on the initiation, progression and disease outcomes may be underestimated. Infection-associated cancers are commonly attributed to viral, and to a lesser extent, parasitic and bacterial etiologies. There is growing evidence that microbial community variation rather than single agent, can influence cancer development, progression, response to therapy, and outcome. We wanted to examine the microbial within sample diversity of paired tumor and adjacent normal tissue across infection-associated cancer types in order to provide an improved understanding of microbial diversity and abundance patterns of the tumor microenvironment and their influence on clinical presentation and survival. We hypothesized that microbial diversity can be a predictor of overall survival and identify racial-related differences that in turn may influence therapeutic decisions.

Methods: We evaluated a subset of tumors in The Cancer genome Atlas (TCGA) from head & neck squamous cell carcinoma (HNSC), liver hepatocellular carcinoma (LIHC) and stomach adenocarcinoma (STAD). Alpha diversity (within sample diversity) as well as presence of specific relevant bacterial and viral species was compared between paired tumor and adjacent normal samples. Association of specific microbial presence with overall survival was also evaluated. Cox proportional regression and generalized linear models were used to correlate bacterial diversity with clinical presentation and overall survival.

Results: A total of 470 paired tumor and adjacent normal were analyzed. In STAD, presence of HHV-4 was associated with poorer survival in the presence of *Selenomonas sputigena* (HR: 2.23, 95% CI 1.26, 3.94, p=0.006) and high diversity index was associated with poorer overall survival outcomes (HR: 2.31, 95%CI 1.1, 4.9, p=0.03). In LIHC, lower microbial diversity were associated with poorer overall survival (HR: 2.57, 95%CI: 1.2, 5.5, p=0.14) while high diversity appeared to have a favorable effect. In HNSC there was a small non-significant trend between high diversity and favorable survival while interestingly, *Rothia mucilaginosa* status was associated with poorer overall survival outcomes (HR: 6.17, 95%CI: 1.3, 29.7, p=0.02 (controlling for smoke)). *Rothia mucilaginosa* was correlated with presence of HPV.

Conclusion: We show a comprehensive analysis of within sample diversity derived directly from human tumor sequences and survival outcomes. Bacterial within sample diversity correlates with survival in HNSC, LIHC and STAD cancers and these associations are dependent on sex and race.

3.2.2 **Introduction**

Microbiological infections account for up to 25% of the total global cancer burden, one of the leading causes of morbidity and mortality worldwide (WHO 2018). Despite declining incidence rates in the United States, cancer remains the leading cause of death among Asian Americans, Native Hawaiians and persons of Hispanic origin (Siegel et al. 2015, Torre et al. 2016). Racial-disparities persist in infection-associated cancers including liver, gastric and cancers of the head & neck (NCI 2015). Liver, gastric and head & neck cancers are in the top ten causes of cancer related deaths worldwide with marked racial differences. Underlying causes for racial disparities are multifactorial and not well understood (Gourin and Podolsky 2006, Singh, Siahpush, and Altekruse 2013, Merchant, Li, and Kim 2014). Recent studies suggest a potential role of microbial composition in racial-related disparities (Brooks et al. 2018). While much effort has gone into the characterization of the gut and oral microbiota, compositional differences of tumor tissue are less explored. Identification of tissue-associated microbial differences is challenging and computationally intensive. There is growing evidence suggesting microbial communities have a dual modulating effect in cancer pathogenesis (Mager 2006, Paulos et al. 2007). In addition, viral-bacterial co-occurrence have been identified to modulate tumor aggressiveness (Pandya et al. 2015). Based on epidemiological and geographic correlations it is suggested that viral agents may interact with specific bacteria resulting in more aggressive tumors and poorer outcomes. For instance, it is recognized that HHV-4 infected stomach tumors are molecularly distinct, while its interaction with *Helicobacter pylori* remains unconfirmed. In hepatocellular carcinoma co-infection with HBV with HCV and their interaction between proteins can also lead to more aggressive tumors. New evidence has hinted at the association of gut microbial dysbiosis with cancer clinical outcomes, even potentially influencing racial-related differences (Gopalakrishnan, Spencer, et al. 2018, Farhana et al. 2018). Gopalakrishnan et al. (2018) determined that a highly diverse gut microbiome can provide improved antitumor response while low diversity with high abundance of unfavorable bacteria such as Bacteroidales, may result in weakened antitumor presentation capacity (Gopalakrishnan, Spencer, et al. 2018). To our knowledge, no studies have examined microbial within sample diversity derived directly from the tumor microenvironment in humans and its relation to survival and potential impact on racial differences. We aimed to study differences in tumor and normal tissue microbiome diversity and determine if this has any relationship with overall survival among infection-associated cancers.

3.2.3 **Material and Methods**

3.2.3.1 TCGA Data: We had previously derived microbial relative abundances and bacterial diversity data from solid tumors from The Cancer Genome Atlas (TCGA) network using bioinformatics workflow based on PathoScope 2.0. Microbial profiles and diversity metrics are available for 22 cancer cohorts. Microbial profiles have associated de-identified relevant clinical data downloaded from data commons under project Project-14778 (Y. Deng, PI). For this work, we selected three infection-associated cancers (head & neck, liver and stomach) to retrospectively examined the relationship between bacterial diversity (within sample diversity) and cancer overall survival. Other caners commonly attributed to infectious etiology were not selected on the basis of low paired sample availability. To determine if a relationship exists, we compared within sample diversity to survival time in tumor and adjacent normal pairs. Bacterial diversity associations to clinical features including basic demographics (age at diagnosis, sex, race and ethnicity), tumor stage, tumor grade, site of resection, histopathology, exposure (alcohol and smoke when available) and viral infection status were also examined.

3.2.3.2 Data availability: The microbial profile data used in this study is maintained by the University of Hawaii Bioinformatics Core, CIM John A Burns School of Medicine per Data Management Plan and User Certification Agreements for each TCGA cohort.

3.2.4 **Statistical Analyses**

Association to cancer overall survival was derived from diversity data among living and deceased groups. To determine the association between difference in bacterial diversity (within sample), in tumor and its adjacent normal samples and the clinical features, pairwise t-test and analysis of variance (ANOVA) were used for multi-group comparisons. Chi-square test were used for categorical data. Using bacterial diversity and significant clinical features as predictors, Cox proportional hazards regression analyses were used to evaluate the associations between diversity and overall survival per cancer type. Tukey's HSD post hoc tests were also carried out. All analyses were carried out n R platform. A complete list of packages and applications is provided as supplemental data.

3.2.5 **Results**

3.2.5.1 Microbial diversity profiles.

Bacterial within sample diversity measured by Shannon-Wiener index of diversity were obtained from a subset of 3 infection-associated cancers from TCGA cancer cohorts per previous (3.1). A total of 470 paired tumor-adjacent normal samples encompassing 235 cases from cancers of the head & neck squamous cell cancers (HNSC), liver hepatocellular carcinoma (LIHC), and stomach adenocarcinoma (STAD), were examined. Of the 470 samples, 11% (1 from HNSC, 28 samples from LIHC and 23 from STAD) had no detectable bacteria presence. 7% of samples, (7 from HNSC, 15 from LIHC and 11 from

STAD) had single bacteria. In both cases diversity indices are represented as zero. All cases with at least one bacteria species present in either tumor or normal were considered for analyses. Viral presence (HBV, HHV-4 or HPV) was detected in 31% of samples without bacterial presence. Overall sample population and microbial profile characteristics and are summarized in **Table 18.**

Table 18 *Microbial diversity profiles among infection associated cancers*

	HNSC		LIHC		STAD	
	Primary Tumor (n=69)	**Solid Tissue Normal (n=69)**	**Primary Tumor (n=81)**	**Solid Tissue Normal (n=81)**	**Primary Tumor (n=85)**	**Solid Tissue Normal (n=85)**
Sex						
FEMALE	21 (30.4%)	21 (30.4%)	35 (43.2%)	35 (43.2%)	37 (43.5%)	37 (43.5%)
MALE	48 (69.6%)	48 (69.6%)	46 (56.8%)	46 (56.8%)	48 (56.5%)	48 (56.5%)
Race						
ASIAN	1 (1.4%)	1 (1.4%)	7 (8.6%)	7 (8.6%)	16 (18.8%)	16 (18.8%)
BLACK	9 (13.0%)	9 (13.0%)	6 (7.4%)	6 (7.4%)	3 (3.5%)	3 (3.5%)
Not reported	3 (4.3%)	3 (4.3%)	4 (4.9%)	4 (4.9%)	12 (14.1%)	12 (14.1%)
WHITE	56 (81.2%)	56 (81.2%)	64 (79.0%)	64 (79.0%)	54 (63.5%)	54 (63.5%)
Age						
Mean (SD)	62.7 (12.2)	62.7 (12.2)	64.3 (14.7)	64.3 (14.7)	67.7 (10.5)	67.7 (10.5)
Median [Min, Max]	62.7 [26.2, 87.7]	62.7 [26.2, 87.7]	67.4 [20.2, 86.0]	67.4 [20.2, 86.0]	69.3 [42.0, 88.5]	69.3 [42.0, 88.5]
Missing	0 (0%)	0 (0%)	2 (2.5%)	2 (2.5%)	0 (0%)	0 (0%)
Vital_Status						
DECEASED	49 (71.0%)	49 (71.0%)	48 (59.3%)	48 (59.3%)	34 (40.0%)	34 (40.0%)
LIVING	20 (29.0%)	20 (29.0%)	33 (40.7%)	33 (40.7%)	51 (60.0%)	51 (60.0%)
HBV status						
Absent	69 (100%)	69 (100%)	75 (92.6%)	75 (92.6%)	85 (100%)	85 (100%)
Present	0 (0%)	0 (0%)	6 (7.4%)	6 (7.4%)	0 (0%)	0 (0%)
EBV status						
Absent	63 (91.3%)	66 (95.7%)	80 (98.8%)	80 (98.8%)	60 (70.6%)	59 (69.4%)
Present	6 (8.7%)	3 (4.3%)	1 (1.2%)	1 (1.2%)	25 (29.4%)	26 (30.6%)
HPV status						
Absent	63 (91.3%)	67 (97.1%)	81 (100%)	81 (100%)	85 (100%)	85 (100%)
Present	6 (8.7%)	2 (2.9%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

	HNSC		LIHC		STAD	
	Primary Tumor (n=69)	**Solid Tissue Normal (n=69)**	**Primary Tumor (n=81)**	**Solid Tissue Normal (n=81)**	**Primary Tumor (n=85)**	**Solid Tissue Normal (n=85)**
Shannon						
Mean (SD)	1.76 (0.863)	1.96 (0.869)	1.23 (1.09)	1.08 (1.04)	1.67 (1.06)	1.57 (1.12)
Median [Min, Max]	1.75 [0.00, 4.09]	2.01 [0.00, 4.04]	1.10 [0.00, 3.21]	0.868 [0.00, 3.03]	1.61 [0.00, 3.69]	1.73 [0.00, 3.55]
Richness						
Mean (SD)	13.6 (21.3)	15.2 (19.8)	11.9 (13.0)	10.3 (11.2)	20.0 (30.2)	15.0 (20.3)
Median [Min, Max]	7.00 [1.00, 114]	9.00 [0.00, 91.0]	9.00 [0.00, 74.0]	7.00 [0.00, 49.0]	7.00 [0.00, 168]	7.00 [0.00, 126]
Evenness						
Mean (SD)	0.907 (0.134)	0.909 (0.129)	0.649 (0.313)	0.624 (0.277)	0.860 (0.176)	0.842 (0.223)
Median [Min, Max]	0.970 [0.412, 1.00]	0.960 [0.310, 1.00]	0.760 [0.0691, 1.00]	0.712 [0.0752, 1.00]	0.934 [0.326, 1.00]	0.958 [0.0478, 1.00]
Missing	4 (5.8%)	4 (5.8%)	19 (23.5%)	24 (29.6%)	16 (18.8%)	18 (21.2%)

Table of population characteristics for HNSC, LIHC and STAD cohorts stratified by sample type. A proportion of samples had single or no bacteria. Cases with no identifiable bacteria in either sample were not considered.

3.2.5.2 Relative abundance differs by race

Gut microbiota composition and relative abundance signatures have been associated with racial differences in colorectal cancer patients and healthy individuals (Farhana et al. 2018, Brooks et al. 2018). We wanted to evaluate if similar patterns existed when examining the microbial composition within the tumor microenvironment. To do this we compared relative abundance profiles stratifying by race.

Figure 21 *Relative abundance in HNSC cohort in tumor and adjacent normal by racial groups*

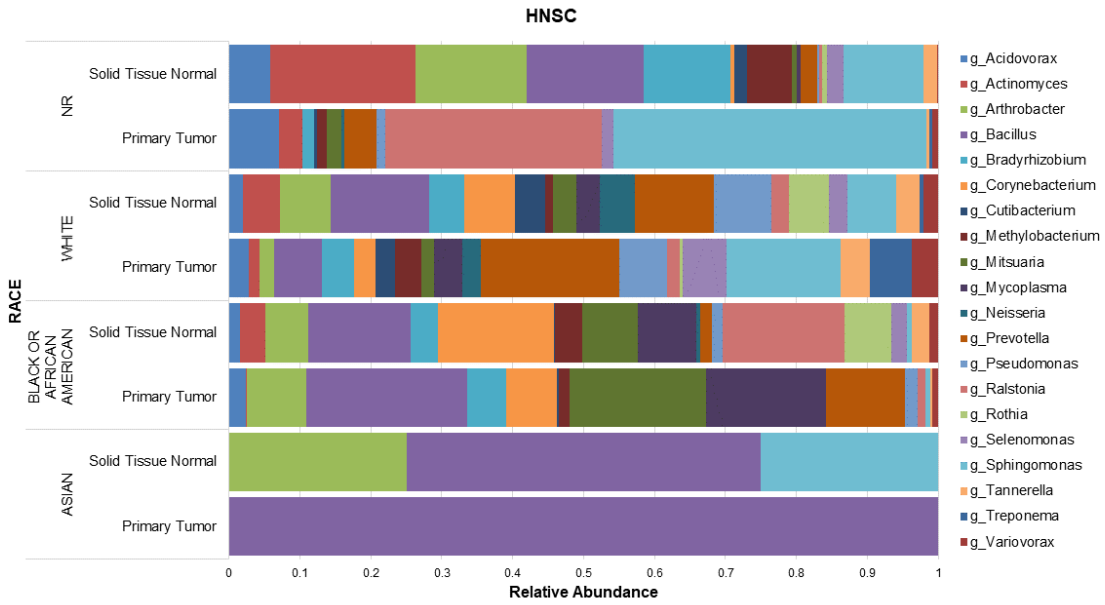


Figure shows top-20 taxa at genus level average relative abundance per sample type across different racial groups within the HNSC cancer cohort. NR: not reported. Differences between tumor and adjacent normal in each racial group are observed. Paired samples of Asian (n=1) and Black or African American (n=9) have less diverse microbial signature patterns compared to White and not reported (considered a mixed-race group). While the tumor samples of White and non-reported have a higher number of different genus, in Asian and Black or African American tumor tissue has far fewer species compared to their paired adjacent normal and to other groups. *Bacillus* spp. (dark purple) are over abundant in the tumor tissue of Asian and Black or African Americans compared to White, where Bacillus is higher in the adjacent normal tissue.

Figure 22 *Relative abundance in LIHC cohort in tumor and adjacent normal by racial groups*

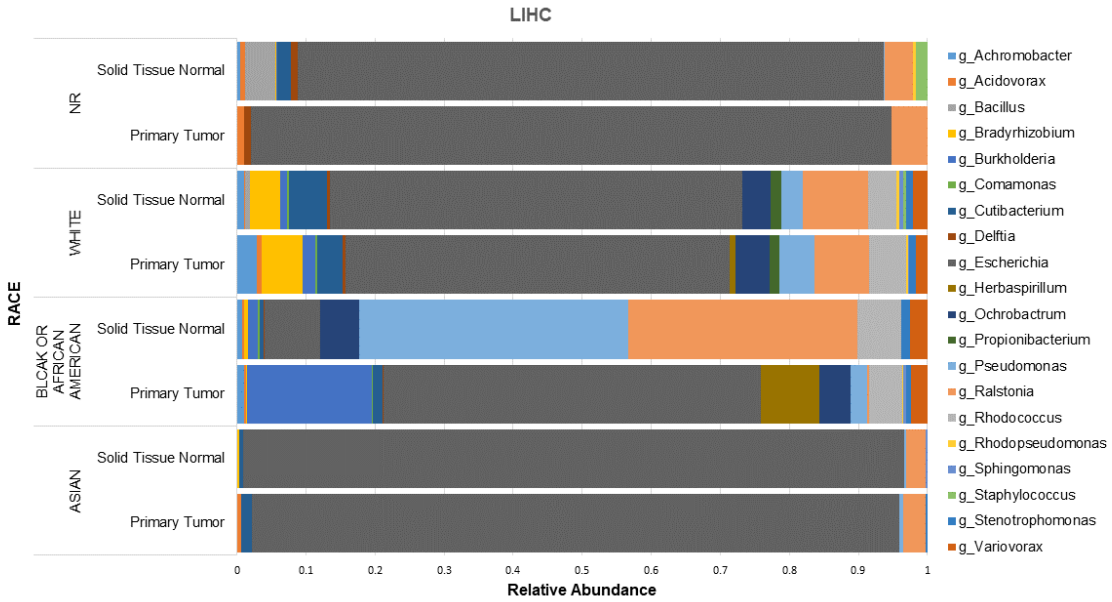


Figure 23 *Relative abundance in STAD cohort in tumor and adjacent normal by racial groups*

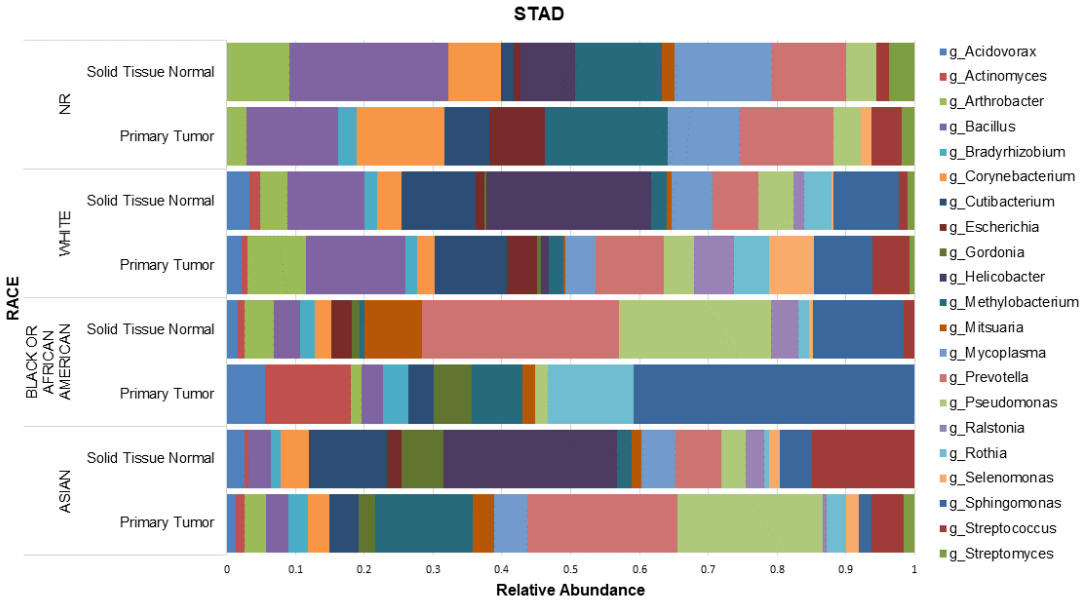


Figure 22 and 23 show top-20 taxa at genus level average relative abundance per sample type across different racial groups within the LIHC and STAD cancer cohorts. NR: not reported. An interaction between submitter site and race was observed in LIHC. In STAD, few species were associated with submitter site which could be indicative of contaminants.

In HNSC, at genus level average relative abundance per sample type across differs by racial groups. *Acidovorax* are present in all except Asian patients (n=1). *Bacillus* are the being the only genus present in the tumor (light purple bar) among Asian. In head & neck, *Sphingomonas* genus is present across all racial groups with the smallest proportion among Black or African American samples. While the tumor samples of White and non-reported appear to have a more diverse tumor microenvironment, in Asian and Black or African American tumor tissue has far fewer species compared to their paired adjacent normal and to other racial groups. Among the three cohorts, LIHC patient samples had the lowest number of taxa at the genus level, with most reads mapped to *Escherichia coli.* Relative abundance patterns are similar across all racial groups paired samples within the cohort. Black or African American group in LIHC shows greater abundance of *Pseudomonas* and *Ralstonia* species. LIHC shares some species with HNSC in similar patterns across Black or African American and White patient samples such as presence of Variovorax which is not present in STAD.

In previous studies (Tae et al. 2014) comparing tumor and matched blood specimens, it was reported that microbial signatures (including virus, bacteria and other species) related ethnic differences where in fact due to specific stamp signatures corresponding to the institutions where specimens originated or where processed. Authors concluded that these signatures could correspond to contamination or laboratory artifacts (Tae et al. 2014).Based on this report, we wanted to know if observed racial differences were associated with the institution from which they originated. Because our strictly paired analyses, if microbial signatures are potentially contamination, we would expect the same relative abundance patterns in tumor and adjacent normal paired samples. We compared relative abundance of specific taxa found to be prevalent and differentially present within the sample population across submitter site institutions and between samples from different racial groups originating from same institution. All HNSC specimens in our study originated and were processed at the same institution and no comparisons were performed. It is presumed that observed variation is attributed to racial and clinical presentation differences. LIHC samples used in this study originated from a diverse population enrolled at different institutions including University of North Carolina, University of Pittsburgh, Mayo Clinic Rochester, Christiana Health, and University of Florida among others. All LIHC specimens were processed at the same institution. In LIHC there were no overall differences across submitter sites in pairwise comparison using Wilcox test with Bonferroni correction for multiple testing. In analyses of variance adjusting for race and sample type, some species were identified as significantly associated with submitter site. *Ralstonia pickettii* ($F(10, 147)= 24.4$, $p=<0$), *Klebsiella pneumoniae* ($F(10, 147)=1.97$, $p=0.04$), *Rhodococcus erythropolis* ($F(10,147)= 2.73$, $p=0.004$) and *Bradyrhizobium japonicum* ($F(10,147)=2.23$, $p= 0.02$). *Ralstonia pickettii* and *Rhodococcus erythropolis* association were dependent on sample type present at different frequencies. Escherichia coli, the most abundant species detected in LIHC cohort, was also associated with submitter site in linear regression model. White specimens (n=1) from a single site, had

greater abundance of Escherichia coli compared to those originating from other sites. A two-way ANOVA was carried out on Escherichia coli relative abundance by submitter site and race. There was a statistically significant interaction between submitter site and race on the relative abundance of Escherichia coli (F(7, 141)=7.5, p=0). A Tukey's HSD post hoc test was carried out. Originating from the University of Florida relative abundance was For Whites compared to Black or African Americans, relative abundance of Escherichia coli was significantly different across institutions (mean difference 380.8, p.adj (BH)=0.04).

STAD, subset of samples originated from different institutions including, Asterand, Christiana Health, the International Genomics Consortium, University of Pittsburgh, ABS-AUPUI, National Cancer Center of Korea, and Indivumed. All STAD specimens in this study were processed at a single institution and different from those in LIHC. However, racial minority groups were uniquely recruited from specific institutions; we presumed that institutions with only one racial group will be inherently different when examining racial differences between institutions. Similarly, we compared the relative abundance of top-taxa within the background (White) population across institutions, and across the racial groups within same institution as applicable. In Asterad, White population specimens had similar abundance patterns of top-taxa in tumor and adjacent normal. Relative abundance differences were observed for *Helicobacter pylori, Streptococcus* and *Selenomonas* species within the White population. Abundance of these species was not associated with submitter site (F(8, 161)=0.6969, p=0.69). Presence of Bacillus subtilis among samples originating from the National Cancer Center of Korea were significantly different compared to other sites (p= 0.006). We found no evidence to suggest interaction between submitter site and race on species relative abundances. When comparing White and Black or African American from the same institution, distinct patterns emerged. Despite, bacterial relative abundance differences between originating institutions, in pairwise comparison with Bonferroni correction were not statistically different. We conclude that observed relative abundance differences between tumor and paired adjacent normal across racial groups in STAD are true observations not due to artifacts from the originating sites. Similar to previous studies gut microbiota findings we find that relative abundances of certain taxa are associated with race.

3.2.5.3 Microbial diversity differs by race

We examined within sample diversity (alpha diversity) from all cases positive for microbial presence in at least one of their tumor or adjacent normal samples. In paired analyses comparing tumor to its adjacent normal, within sample diversity indices were not significantly different for any of the three cohorts (**Figure 24**). Approximately 8% (n=19) of the samples had no reported racial background. We considered these to be a mixed group. We performed student t-test for continuous variables and Chi square test for categorical variables to examine differences in population characteristics in order to ascertain differences in microbial abundance and diversity profiles. Across cohorts, stratifying by sample type, there were no

significant differences in the proportion of male-to-females, the age at diagnosis, vital status. There were significant differences in the proportion of racial minorities as previously noted by Zhang et al. (Zhang et al. 2017) and the presence and absence of viral agents associated with cancer initiation and aggressiveness (Pandya 2015). Microbial profiles show similar patterns in tumor and adjacent tissue samples across different racial groups (**Figure 25**). Figure 25 shows composite bar graph by sample type and racial background across HNSC, LIHC and STAD cancers. Evenness, a measure of the relative abundance is illustrated with a line graph on a secondary axis which show the relation between the diversity index and species abundance from 1 indicating complete evenness (most species are similar) and 0 highly diverse. We also examined clinical presentation and exposures within the groups across the racial groups. Population differences within the cohorts were considered to build proportional hazards models.

In HNSC, we observed small trends in the number of species (p= 0.05). When stratifying by race, stage, vital status, morphology or tumor grade did not differ significantly. Bacterial diversity differed by anatomical site. Among HNSC patients, there were no significant differences, when considering bacterial species relative abundance and presence or absence of in microbial profiles at any level, including *Corynebacterium pseudotuberculosis*, *Rothia mucilaginosa*, *Capnocytophaga* ChDC OS43, *Treponema denticola*, *Ralstonia pickettii*, *Streptomyces gilvosporeus*, *Sphingomonas melonis*, *Sphingomonas sanxanigenens*, and *Arthobacter*. Two *Actinomyces* species, *Actinomyces pacaensis* (presence, not its relative abundance) and *Actinomyces myeri* (its relative abundance, not its presence) were significantly different across the racial groups within HNSC (p=0.03). When stratifying by race, there were no differences in exposures including the number of cigarettes per day; however, significant differences in smoke years were observed (p=0.001). Diversity means by race did not differed significantly in HNSC. (**Table 18**).

For LIHC there were no differences in clinicopathological presentation including primary diagnosis and resection site, except for tumor stage when stratifying by race and sex. In LIHC race proportions were in itself significantly different with no female of Asian background (**Table 18**). While species richness was not significant across the different racial groups in LIHC, within sample diversity measures were significantly different among racial groups (Student-t-test, Shannon Index of diversity, p=0.01; evenness, p=<0.001). We also observed a significant difference in the ratio of males-female patients among racial groups although there were no differences overall. There were no differences in survival status by race, however we noted significant differences in the days survived by racial groups (Student t-test, p=0.03). Asian patients within the LIHC cohort were significantly younger (mean 53.3 (10.5)) compared to White (mean 65.8 (15.2)) and the non-reported mix race group (mean 62.4 (11.9)) (Wilcoxon, p=<0.001). Microbial profiles for LIHC patients were significantly different when stratifying by race and sex. Viral presence of HBV differed both by sex and race while presence of HHV-4 did not (**Table 18**). HBV

presence was significantly higher among Asian patients (presence and abundance p=<0.001). *Bradyrhizobium* spp. presence were slightly different by sex in LIHC. In pairwise tests, diversity was significantly different between White and Asian patients (BH p.adj=0.009). We conclude that differences in bacterial diversity for LIHC differ by race and this relationship is dependent on HHV-4 infection status (F(1, 53)=4.8 p=0.03).

Table 19 *Shannon-Weiner diversity index in tumor and adjacent normal pairs*

	Primary Tumor		Adjacent Normal		
Cohort	Female	Male	Female	Male	pval
HNSC					
White	1.75 (0.79)	1.68 (0.75)	2.05 (0.95)	1.90 (0.82)	
Asian	--	0.69 (0.00)	--	1.39 (0.00)	
Black	1.36 (0.03)	2.26 (1.05)	1.78 (0.17)	2.05 (1.08)	
Not reported	0.69 (0.00)	2.97 (1.12)	1.35 (0.00)	2.96 (1.08)	0.36
LIHC					
White	1.28 (1.01)	1.47 (1.10)	1.31 (1.02)	1.13 (1.05)	
Asian	--	0.36 (0.48)	--	0.13 (0.26)	
Black	0.05 (0.05)	1.01 (1.24)	0.00 (0.00)	1.01 (1.07)	
Not Reported	2.02 (1.10)	0.43 (0.43)	0.92 (0.67)	1.26 (0.52)	0.001
STAD					
White	2.06 (0.87)	1.71 (0.97)	1.51 (1.09)	1.73 (1.05)	
Asian	1.28 (1.29)	1.08 (1.33)	1.56 (1.45)	1.19 (1.42)	
Black	--	2.27 (0.68)	--	2.83 (0.25)	
Not Reported	1.35 (0.21)	1.15 (0.87)	1.19 (0.85)	1.39 (0.76)	0.007

Table shows with sample diversity indices across cohorts by different racial groups. Females of Asian background are underrepresented in HNSC and LIHC cohorts and Blacks females in STAD. Analyses of variance examining adjusting for sample type, race and sex shows significant differences in within sample diversity in LIHC and STAD

Significant within sample bacterial diversity were observed in STAD cohort where Black or African American patients fell within the highest diversity quartiles both in tumor and adjacent normal. There were significant differences with histopathological grade, and site of resection while no difference in in tumor staging within the population. When stratifying by race, significant differences in primary diagnosis (p=0.022) and age at diagnosis (p=0.027), and site of resection (p=0.022) were observed. Viral presence was significant among females of different racial groups. HHV-4 positive status was significantly higher among White females compared to other racial groups (13% non-White). Microbial species detected differed across racial groups and by sample type. Several species including *Fusobacterium nucleatum*, *Lactobacillus amylovorrus*, *Lactobacillus salivarius*, *Campylobacter concisus*, *Lactobacillus fermentum*, *Neisseria enlongata* were unique to tumor samples. *Bacillus subtilis* (presence, not its relative abundance), was deferred by race with highest prevalence among White patients with differential

abundance in tumor compared to its adjacent normal (White: 61% versus 59.3%, Not reported 50% versus 66.7%, Black 33% versus 66.7%, and no difference among Asians, (F(4,165)=8.7, p=<0.0). *Cutibacterium acnes, Mycoplams mycoides* and *Ralstonia pickettii* presence were significantly different among different racial groups in tumor, while in adjacent normal, *Arthrobacter* presence was the different among racial groups. However there were differences between females from different racial groups in STAD.

Figure 24 Bacterial within sample diversity across cohorts comparing tumor to its paired adjacent normal tissue

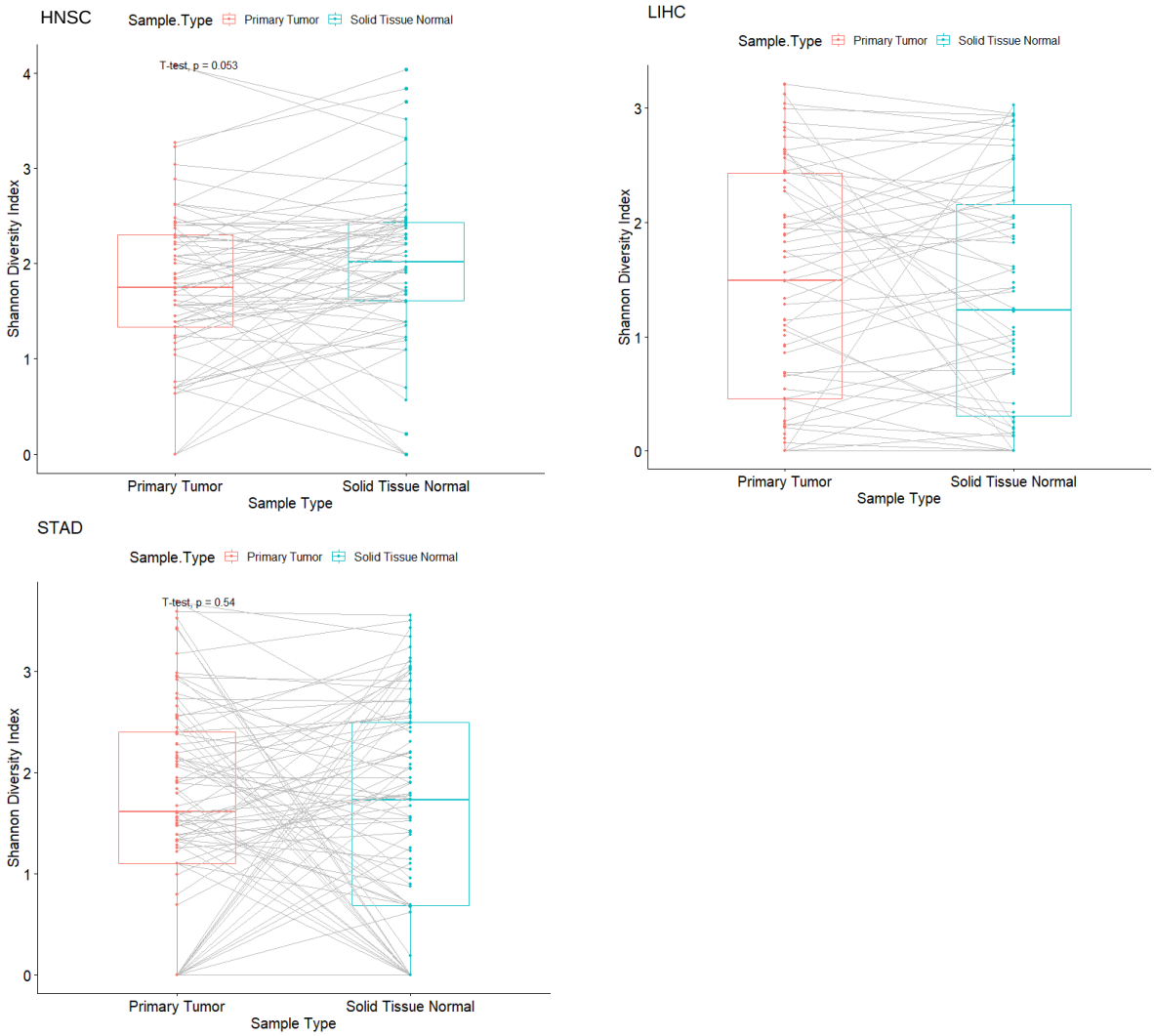


Figure shows paired t-test comparing tumor to adjacent normal within sample bacterial diversity. HNSC has a marginal trend of lower diversity in tumor compared to adjacent normal at 0.05 compared to no difference in LIHC and STAD. Despite there is an observable bimodal trend in STAD with presence and absence of taxa resulting in diversity index of zero in one sample over the other and a slightly higher mean in adjacent normal samples. Compared to LIHC where there is no observable patterns when comparing tumor to its paired adjacent normal.

Figure 25 *Microbial profiles of infection-associated cancers of the head& neck, liver and stomach*

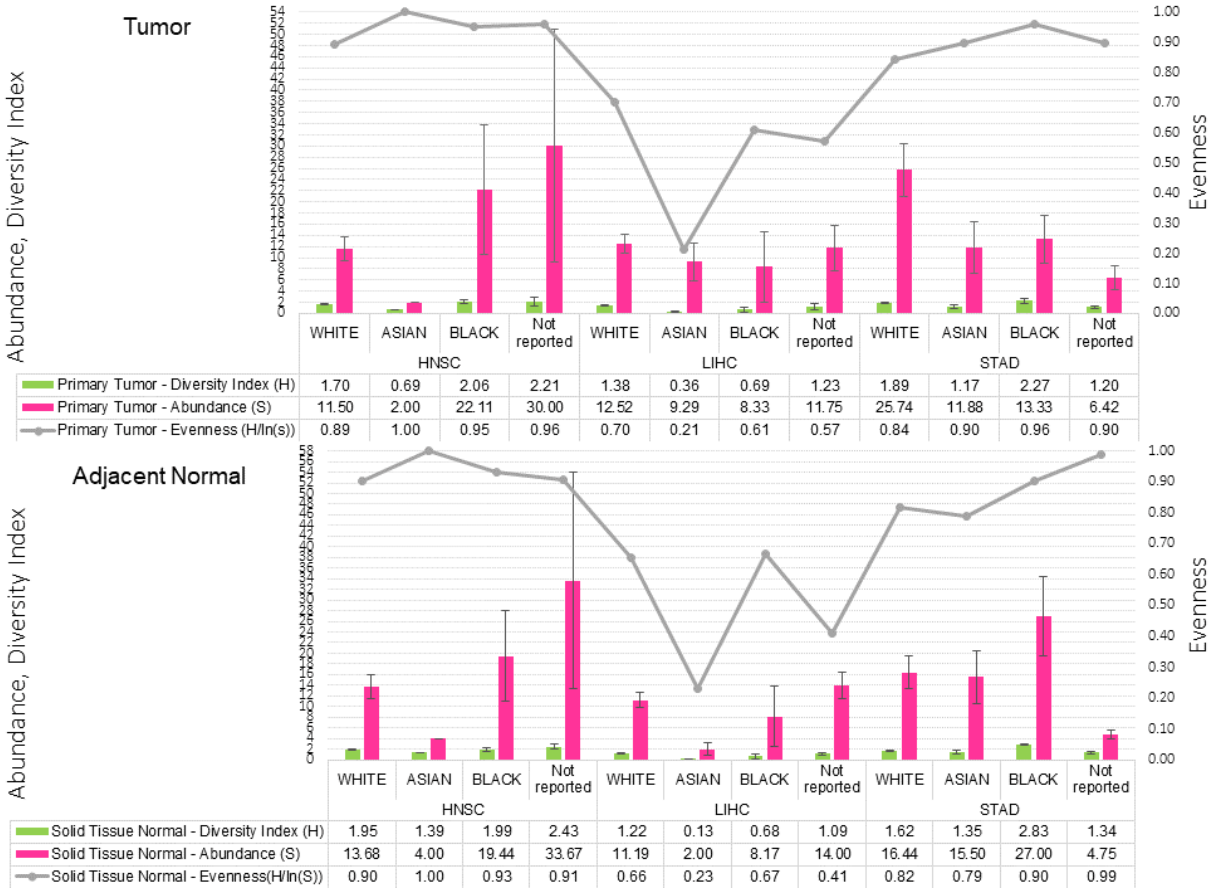


Figure shows the microbial diversity profiles for infection associated cancers in tumor and adjacent tissue samples among different racial groups with standard error bars. Average absolute observed abundance (pink bars), within sample diversity index as determined by Shannon-Weiner index (green bars), and evenness (y-axis) of species abundance per racial group in each cohort is shown as line graph. Shannon diversity index measures the relative proportion of species abundance (Summation of species by the natural log of total species within a sample). Evenness is a measure of the relative abundance (Shannon index by natural log of total species) from 1 indicating complete evenness and 0 highly diverse. Here is being used as a complement to index of diversity calculated with vegan R-package in which both are calculated separately. In HNSC, Asian (n=1) have the lowest combined diversity measures (low number of species with high degree of evenness) compared to Asian in LIHC (n=7) and STAD (n=16) where despite absolute abundance being low, the species present are from diverse origin. White patients in contrast, although number of species within the sample are high, these appear to be from similar origin resulting in low combined indices of diversity. Overall although similar, diversity indices between tumor and adjacent normal are contrasted specially in LIHC and STAD cohorts.

Figure 26 Racial diversity differences by cohort in tumor and adjacent normal samples

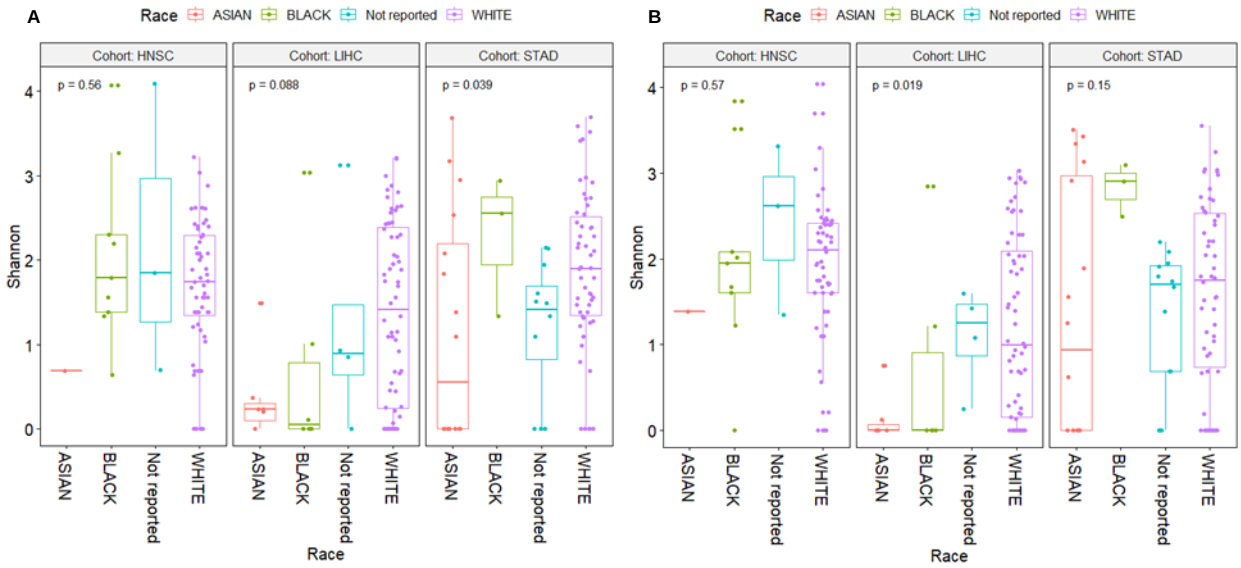
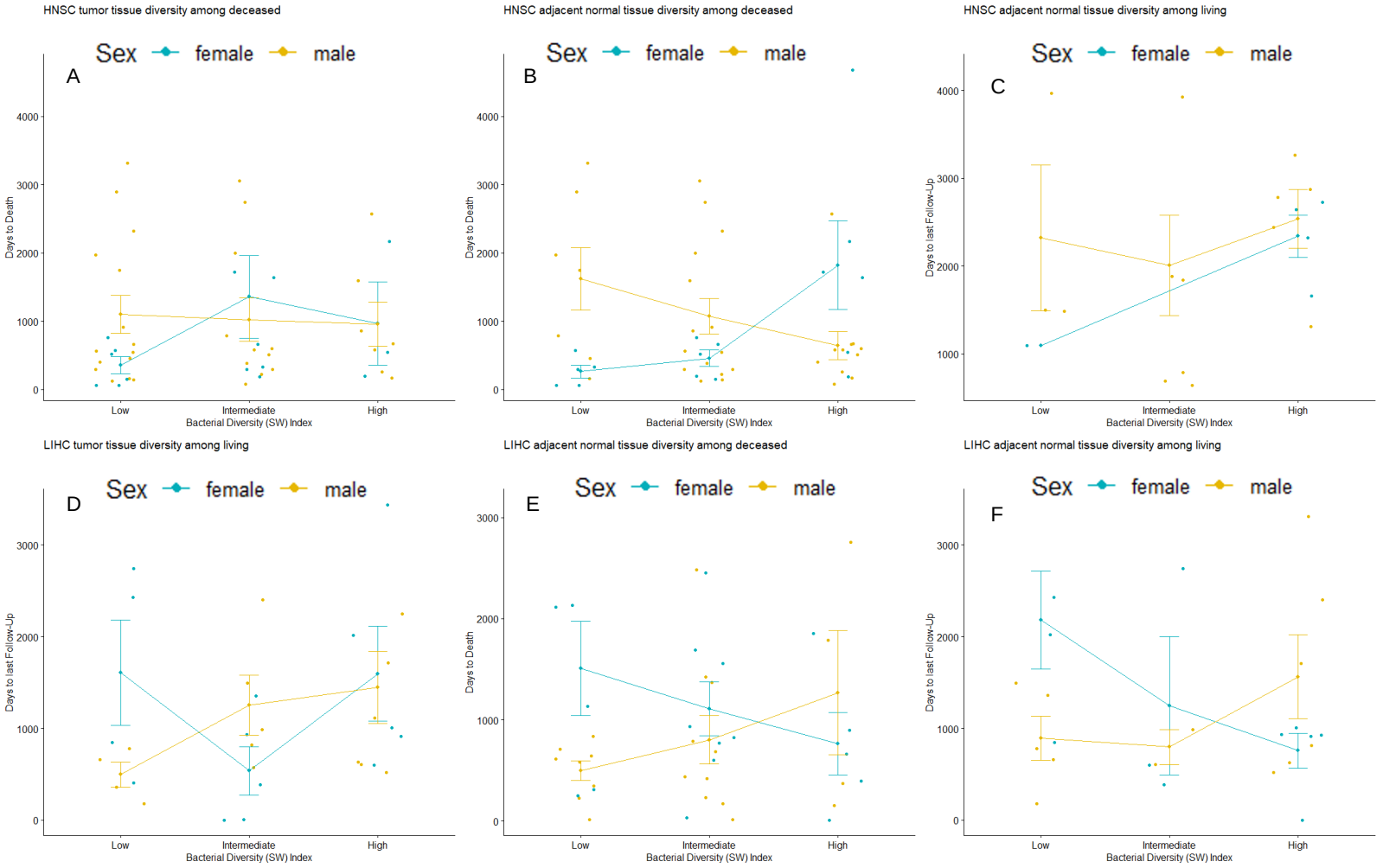


Figure shows Wilcoxon (unpaired) test comparing Shannon bacterial diversity index by racial groups across cohorts in tumor (A) and adjacent normal (B). Bacterial within sample diversity is significantly different in STAD tumor (Wilcoxon, p=0.4) and in LIHC adjacent normal (Wilcoxon, p=0.02), while no difference is observed in HNSC in either tissue type among racial groups.

3.2.5.4 Microbial within sample diversity is associated with overall survival.

Based on the different patterns observed in bacterial within sample diversity identified in each cohort, we wanted to test for the interaction between diversity and race, sex and other clinical features to overall survival. Comparisons were completed by vital status. Overall, among HNSC patients, there were no significant associations between overall survival and clinicopathological features including sex and microbial diversity. Yet significant interaction was observed between sex and microbial diversity of the adjacent tissue relationship to survival when stratifying by vital status (two-way ANOVA, F(2, 43)=6.28 p=0.004)). We then tested the relationship of diversity and overall survival stratifying based on tertiles of Shannon Diversity index. HNSC deceased male patients with high bacterial diversity in the adjacent normal tissue, on average lived shorter days after diagnosis than males with low diversity. While female counterparts had opposite effects (**Figure 27,** B). There were no significant associations in HNSC between diversity quartiles and survival. We conclude that any relationship between diversity and survival are dependent on the interaction with sex in HNSC cohort subset. Although not significant, HPV status appears to have an effect on survival which is dependent in bacterial presence and smoking status.

Figure 27 Bacterial Diversity relationship with survival in HNSC and LIHC is dependent on sex



Line plots with mean and standard deviation of the association interaction between sex and bacterial diversity ranges.

Rothia mucilaginosa was found to be differentially abundant, predominant in adjacent tissue (log2fc=-4.44, p=0.001 FDR=0.32). In Cox proportional hazards, we observed a small a non-significant association with detrimental and HPV positive infection status (**Figure 28**).

Figure 28 *HNSC Hazard Ratios based on Cox proportional hazards*

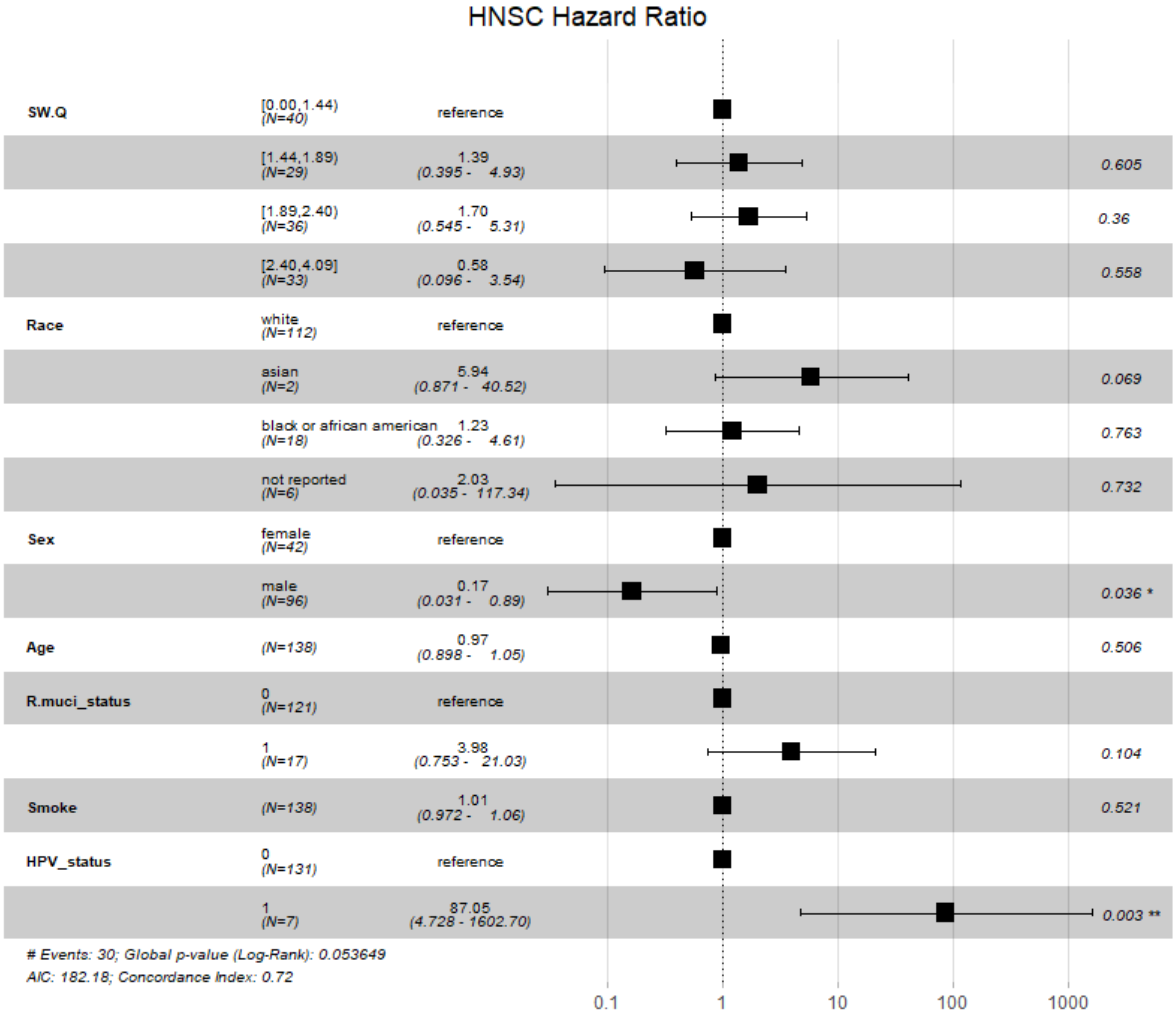


Figure shows Cox proportional hazards forest plot based on adjusted model with interaction terms of smoke and clinicopathological characteristics within the HNSC cohort. Overall, there is no increased risk associated with bacterial diversity. Males are at lower risk than females and HPV status appears to have an association with detrimental survival while bacterial presence of R mucilaginosa has no affect.

In LIHC bacterial diversity was associated with overall survival with significant differences between low to intermediate levels diversities compared to low and high diversity index quartiles (**Figure 29**). Sex, race

Figure 29 LIHC overall survival is associated with microbial within sample diversity

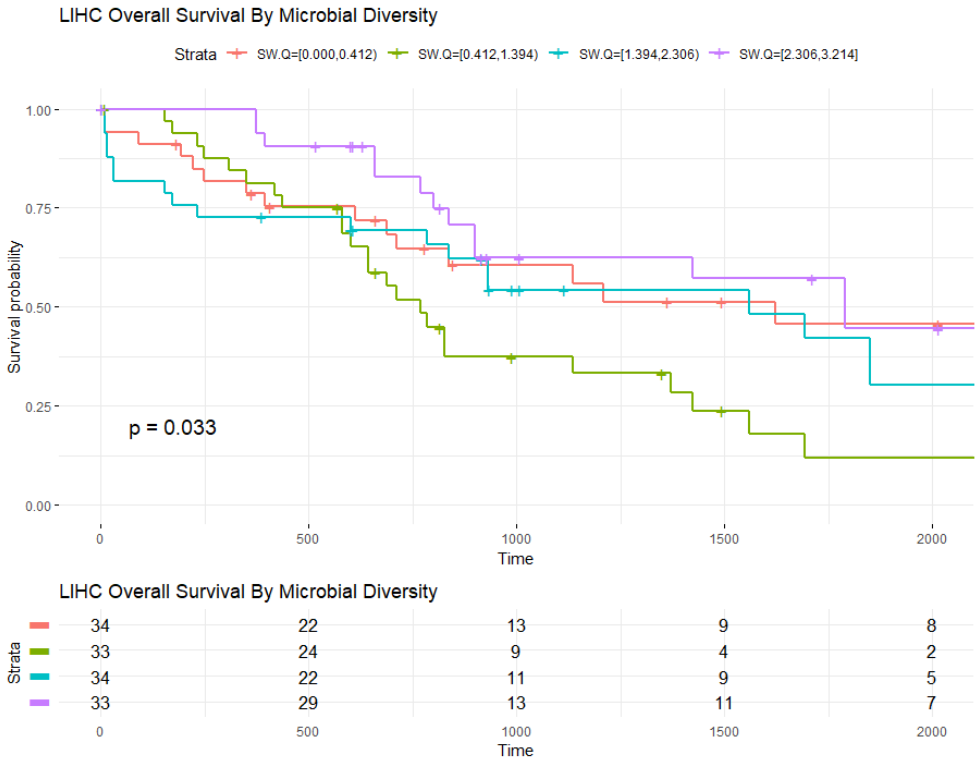


Figure shows survival curve based on Kaplan Meier estimates (created with survminer) for the relationship between bacterial within sample diversity and overall survival with global pvalue. Strata with risk set size is shown in table. Time in days survived after diagnosis or days to last follow up (censored). Censoring over time are delineated by "+" within the survival lines for each strata. Low diversity =0 to 0.4, intermediate-low =.4 to 1.4, intermediate-high=1.4 to 2.3, and high diversity =2.3 to 3.2. Intermediate-low versus low diversity (HR: 2.4, 95%CI: 1.2-5.0, p=0.02). Proportional assumptions violations corrected in Cox hazards model

and age were also associated with overall survival in LIHC. Similar to HNSC, LIHC cohort had an interaction between sex and diversity tertiles. Among living males (n=16), lower tumor bacterial diversity is associated with shorter days survived after diagnosis with opposite effect in females (n=14) at the same diversity range. Contrary to HNSC, bacterial diversity in the adjacent tissue had a negative effect among females and favorable effect among males (**Figure 27**, E). Intermediate diversities are associated with poorer survival compared to higher diversity quartiles (Kaplan Myer, p=033, Figure 29).

In Cox logistic regression model, intermediate-low bacterial diversity was associated with more than double the risk compared to those at low, intermediate-high and high diversities(HR: 2.4, 95%CI: 1.7-5.0, p=0.02), as is positive HBV infection status (HR=5.8, 95%CI:1.4-24.7 p=0.02). Analysis of deviance revealed significant interaction between age at diagnosis and HBV infection status with overall survival (supplemental data).

Figure 30 *LIHC Hazard Ratios based on Cox proportional hazards*

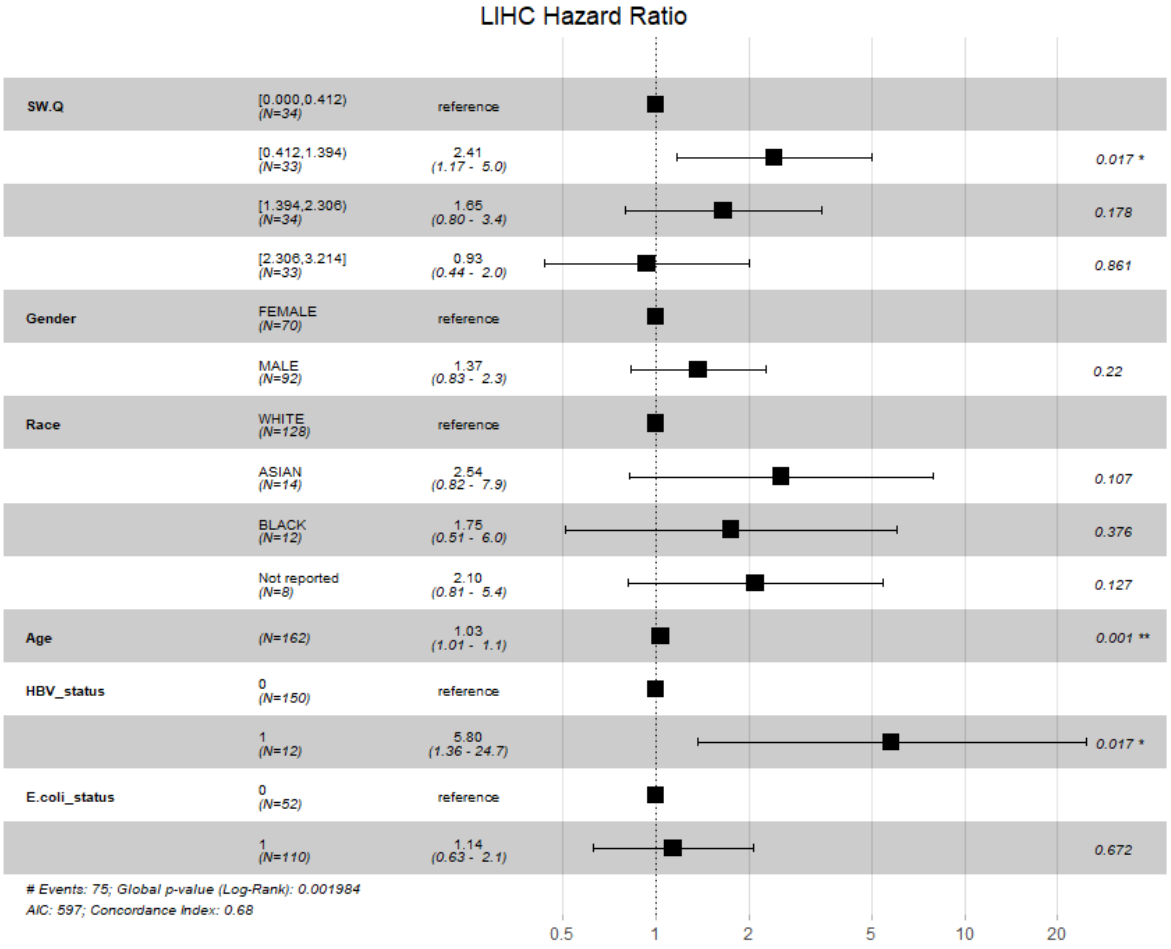
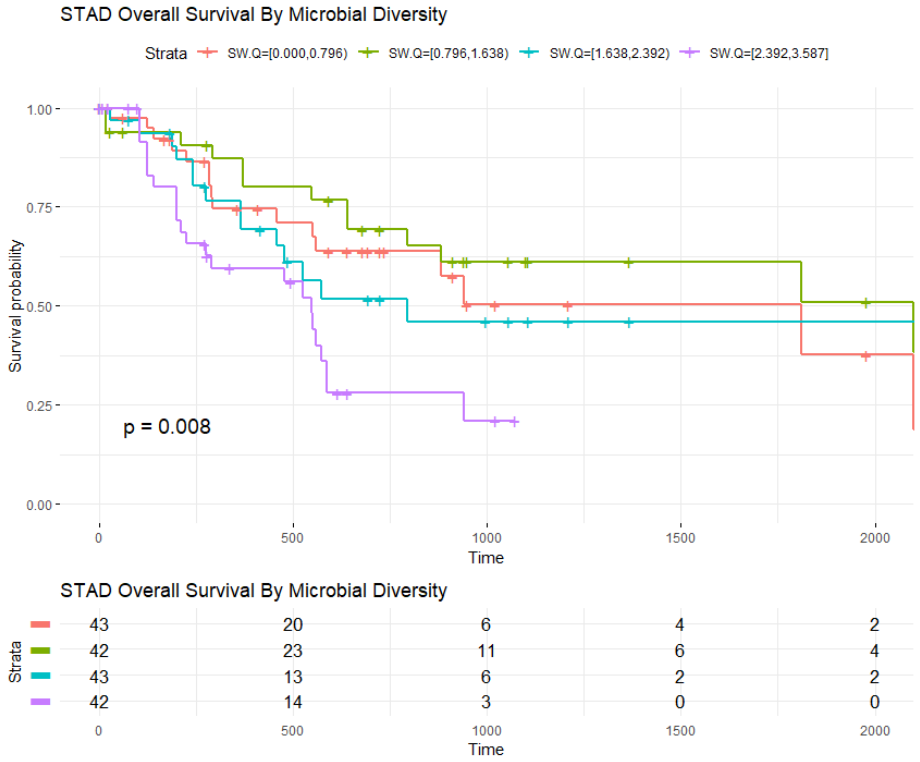


Figure shows Cox proportional hazards forest plot based on adjusted model with interaction of clinicopathological characteristics found to be significantly different within the LIHC cohort. After adjusting for interacting features, diversity intermediate-low quartile are at higher risk.

In STAD, overall survival was associated with bacterial diversity quartiles where higher diversity resulted in poorer survival outcomes (**Figure 31**). Significant interaction between sex and tumor diversity tertiles was observed among those alive at time of censoring ($F(2, 45)=3.6$, $p=0.03$). Post hoc analyses revealed significant difference between males with high diversity compared to males with intermediate diversity indices (diff=-888, -1762 to -13.4, adjusted p value (BH)=0.04). Overall survival was associated with

demographic characteristics including race (p=0.006) and age at diagnosis (p= 0.001). There was no correlation between overall survival and sex alone. We then tested the correlation between diversity, race

Figure 31 STAD overall survival is associated with microbial within sample diversity



STAD Kaplan Meier survival curve (created with survminer) for the relationship between bacterial within sample diversity and overall survival with global pvalue (p=0.008). Strata with risk set size are shown in table. Time in days survived after diagnosis or days to last follow up (censored). Censoring over time are delineated by "+" within the survival lines for each strata. Low diversity =0 to 0.8, intermediate-low =.8 to 1.6, intermediate-high=1.6 to 2.4, and high diversity =2.4 to 3.6. Low versus high diversity (HR: 2.8, 95%CI: 1.3-6.0, p=0.01). Proportional assumptions violations corrected in Cox hazards model (supplemental data).

and age at diagnosis. We observed there was correlation between bacterial diversity and race (Chi-sq =27.4, df=9, p=0.001) and no correlation with age at diagnosis (rho= -0.008, 95%CI= -0.235, 0.063, p=0.25). Analyses of variance were carried out to compare the relationship between survival days and diversity controlling for both age and race. We concluded that the relationship between diversity and survival days is dependent on the interaction with race, where Asian patients with lower diversity have better survival outcomes compared to White (HR=0.16, 95%CI=. 0.37, 0.7, p=0.014). Hazards by clinicopathological features were also examined. In STAD, tumor histopathological stage was associated with diversity. Bacterial diversity was on average higher at higher tumor stages. Whites classified at tumor stage III had on average higher diversity indices than other racial group. Tumor stage was in turn associated to survival days (Chi-sq=43.3, p=<0.0). Relationship between survival days and stage was dependent on diversity (F(27,129)=2.3, p=<0.001). Cox proportional hazards revealed that those

classified at tumor stages II to III with lower diversity indices survived longer days after diagnosis (HR=0.037, 95%CI=0.17, 0.8, p=0.01).

Figure 32 *STAD Hazard Ratios based on Cox proportional hazards*

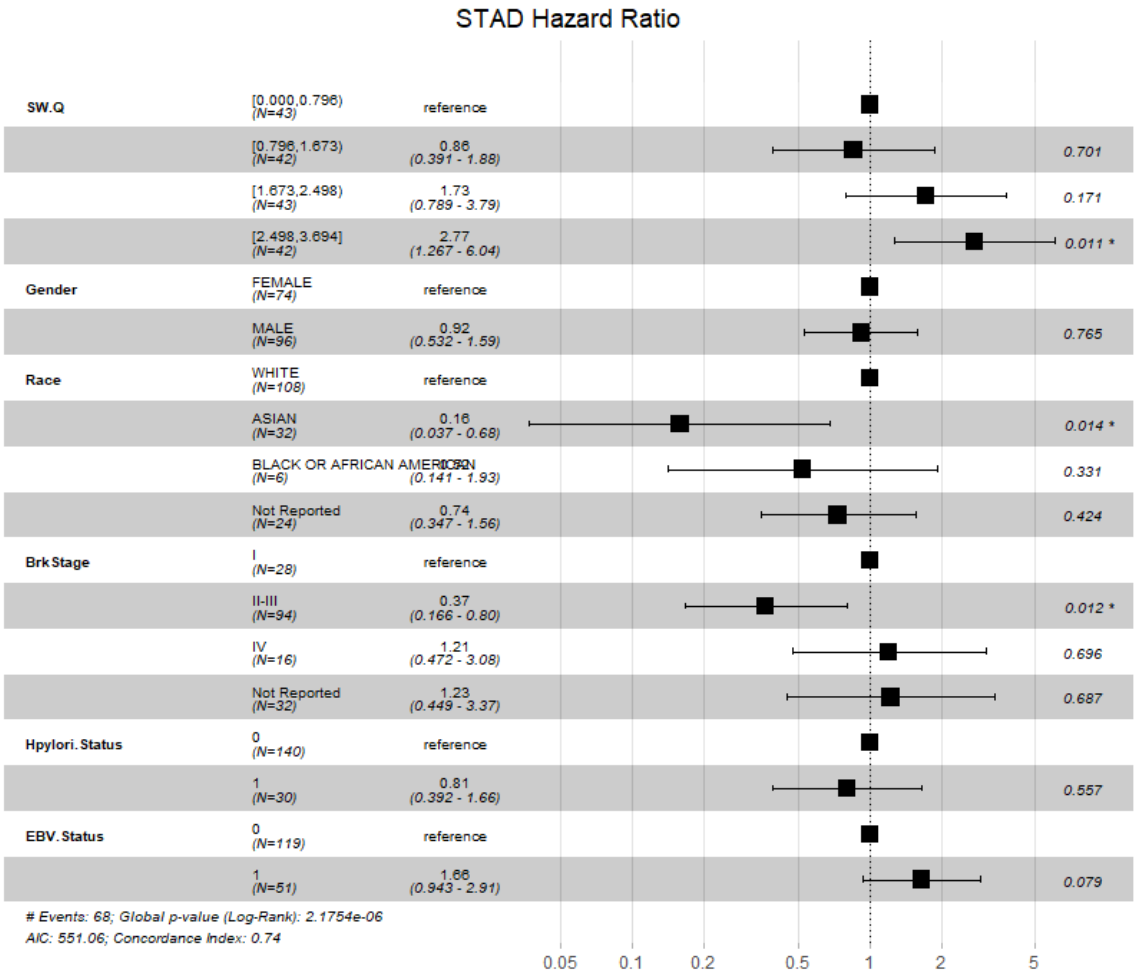


Figure shows Cox proportional hazards forest plot based on adjusted model with clinicopathological characteristics found to be significantly different within the STAD cohort. Infection status have no effect on survival. Increase in diversity is proportionally associted with increased risk. Patients within the higher diversity quartiles are at increased risk (HR=2.8, 95%CI=1.3, 6.0, p=0.01).

3.2.6 **Discussion**

In high throughput sequencing data while microbial cancer associations have gained interest in recent years, insufficient attention has been paid to the potential of addressing racial disparities. Previous studies have identified bacterial diversity as a modulator of treatment response where a highly diverse gut microbiota that includes beneficial bacteria exerts beneficial treatment outcomes; compared to a gut with a less diverse ensemble and high prevalence of pathogenic bacteria which can have the opposite effects (Gopalakrishnan, Spencer, et al. 2018). Other studies have found that gut and oral microbiota diversity contributes to racial differences in cancer (Farhana et al. 2018) and in healthy adults (Brooks et al. 2018, Gupta, Paul, and Dutta 2017, Hoffman et al. 2018). Studies have focused on tumor viral detection or bacterial metagenomics profiling of the gut and oral microbiota. In this study we examined the relationship between bacterial relative abundances and diversity to overall survival in three infection-associated cancers of the stomach, liver and head & neck to determine if similar to previous studies we could identify bacterial differential patterns associated with race and overall survival. To do this we utilized previously created microbial abundance and diversity profiles from a total of 470 paired tumor and adjacent normal samples corresponding to 235 cases. Across the three cohorts, majority, 74% (n=174) were self-reported as non-Hispanic White, 10% were Asian, 8% were Black or African American and 8% were of no reported racial background and considered to be a mixed race group. From these, 40% (n=93) were females, and 55% (n=131) were deceased. We compared bacterial diversity associations to clinical features including basic demographics (age at diagnosis, sex, race and ethnicity), tumor stage, tumor grade, site of resection and histopathology. Smoke exposure (in head & neck cancers) and viral infection status of Epstein-Barr virus (HHV-4), human herpes virus –B (HBV) and human papilloma virus (HPV) were also examined. We found that across all three infection-associated cancers examined here, the relationship between microbial diversity and survival differs by tissue type, tumor versus the adjacent normal and are dependent on the interactions with sex and race. Interactions with race and sex also differ by survival status. We believe that the interactions with sex are perhaps related to the absence lower number or absence of females within the 3 cohorts examined, these associations are stronger in gastric and liver cancers than in cancers of the head & neck. We found that microbial tumor and adjacent tissue diversity were on average lower among persons of Asian background compared to White counterparts. African Americas in the other hand, had similar within sample diversity to White in HNSC, lower in LIHC and higher in STAD. We also noted that bacterial within sample diversity relationship to overall survival had opposite effects in STAD compared to LIHC while no differences were observed in HNSC.

In STAD cohort, higher bacterial diversity was associated with poorer outcomes (HR= 2.77, 95%CI=1.3, 6, p=0.01) with differences among different racial groups. For example Asian persons had lower risk of death compared to White (HR=0.16, 95%CI=0.04, 0.7, p=0.01). Yet, among those who were deceased,

tumor average bacterial diversity was higher among Asian compared to other groups. Among those living at time of censoring, there was a divergent association with days survived after diagnosis. Here, higher tumor tissue within sample diversity, appears to be associated with longer days, while higher diversity in the adjacent tissue is associated shorter days. We noted that Asian patients had the lowest average diversity indices in the adjacent tissue. This could explain the protective effect Asian persons had within our STAD subset compared to White. Black or African Americans (n=3) had higher bacterial diversity in the adjacent normal which similarly could be associated with their poorer outcomes. The bimodal diversity is suggestive of microbial species overturn (dysbiosis) and colonization in disease progression. We also examined microbial (viral and bacterial) presence, relative abundance and co-occurrence patterns. In STAD, presence of HHV-4 was associated with poorer survival (HR: 2.23, 95% CI 1.26, 3.94, p=0.006). However relative abundance of HHV-4 was not correlated with survival days. Prevalence of HHV-4 was higher among non-Hispanic Whites who suffered the poorer survival outcomes in the STAD subset. In LIHC, lower microbial diversity were associated with poorer overall survival (HR: 2.57, 95%CI: 1.2, 5.5, p=0.14) while high diversity appeared to have a favorable effect overall. Interestingly we observed that compared to White, Asian and African American patients who were deceased at censoring, had low bacterial diversities and opposite outcomes were Asian appear to do poorer while African American have longer survival days. Bacterial diversity in the adjacent normal appears to have no effect while diversity in the tumor does. Among deceased LIHC patients, bacterial within sample diversity in the tumor is lower diversity is associated with lower number of days survived after diagnosis. We also found a divergent effect of bacterial presence and relative abundance on race and sex and tumor stage. Interactions between several clinicopathological features and diversity were found. We tested for the possibility of site submitter interaction and found that cohorts have specific signatures by race which are associated or dependent on age and sex. *Escherichia coli* presence and relative abundance in LIHC of was associated with submitter site and race. In linear regression model accounting for interacting terms, submitter site was associated with race. In analyses of variance this explained 27% of the variation (R^2=0.27, adjuster - R^2-squared =.20. In pairwise analyses using Wilcoxon rank sum test, sites with similar racial recruitment were highly correlated. We interpreted this as the difference in recruitment for racial minorities across different cohorts were the effect of the site is modulated by race. Since racial groups were recruited from specific sites it is logical that the two are interacting, while White patients were recruited from multiple sites. These interaction effects are predominant in LIHC cohort. We observed that for some bacterial species there was significant interaction between bacterial abundance and the originating institution. Previous studies have suggested that microbial differential patterns are dependent the submitting institution rather than other features including demographics of age, sex and race(Tae et al. 2014). In our analyses we found that in LIHC, bacterial signatures patterns differ by submitting institution with high correlation between sites with similar enrollment patterns. We discovered an interaction between the submitting institution, sample type and race. Although the possible effect of contamination at submitter

cannot be ruled out, our paired analyses design should account for this. Small sample representation of racial minorities may influence our ability to detect hidden or masked associations. Future studies with larger sample size should confirm our findings. Here we shown a comprehensive analysis of within sample diversity and relative abundance derived directly from human tumor sequences and survival outcomes. We conclude that bacterial within sample diversity correlates with survival in HNSC, LIHC and STAD cancers and these associations are dependent on the interactions with sex and race with divergent associations of beneficial or detrimental effects with overall survival by cancer type.

3.2.8 **Literature Cited**

Located in Appendix E (pp 147)

CHAPTER IV CONCLUSIONS & FUTURE DIRECTIONS

4.1 **Summary of conclusions**

In this study we built microbial profiles for solid tumor cancers within the TCGA Network in order to determine epidemiological and clinical significance of microbial differential abundance between paired and adjacent normal tissue using whole exome sequencing data. We screened more than 200 billion whole exome sequencing reads, generating detailed microbial documentation for more than 800 tumor and adjacent normal samples across 9 cancers types. We have presented the microbial composition in STAD, LIHC, COAD, READ, LUSC, LUAD, HNSC, CESC and BLCA cohorts and examined potential roles in carcinogenesis, overall survival, as well as the impact on racial disparities in STAD, LIHC and HNSC cancers. We demonstrated the ability to identify differential composition of bacteria species from human tissue creating a platform for current and future research while addressing important health disparities research questions. This dissertation advances the concepts of cancer tumor microenvironment by the detailed examination of both, the tumor and adjacent tissue microbial composition directly derived from human sequences. We showed that bacterial shifts between tumor and adjacent normal, as well as their microbial diversity and differential abundances could be indicative of an active role of bacteria in disease progression and cancer pathogenesis determined by overall survival outcomes. Further, validation by qPCR of *Helicobacter pylori, Fusobacterium nucleatum and Selenomonas sputigena* in gastric cancer and *Selenomonas sputigena* in lung adenocarcinoma with tissue samples from an independent population from the Hawaii Tumor Registry demonstrates that our simple methods for the identification and interpretation of complex microbial data, are equivalent to complex transcriptomics and metatranscriptomics methods. Altogether this study has provided a cross-cancer view of tumor-bacterial associations.

Our results show that bacterial within sample diversity correlates with overall survival which, confirm and extend current knowledge of bacterial diversity patterns and viral co-occurrence as demonstrated by HPV, HHV-4 and HBV patterns of co-occurrence. Importantly our results argue against monomicrobial actions on cancer pathogenesis and overall survival. Furthermore, it highlights the importance of utilizing paired sample data together with measures of absolute read abundance, relative abundance and population prevalence in studies evaluating microbial composition and their impact on differential outcomes in order to avoid wrongful conclusions. We also highlight the following findings:

- Significant differential abundance of *Helicobacter pylori* in STAD and *Bacteroides vulgatus* in COAD tumor compared to adjacent normal samples. Differential abundance may be indicative of disease progression rather than a beneficial or protective effect. In disease progression, bacterial community may lead to unfavorable effects by destroying the integrity of cellular barrier, enabling pathogenic bacteria to exert damage or reactivating viral pathogens.

- Diversity is correlated with tumor stage and overall survival in HNSC, CESC, LUSC, LUAD, STAD and LIHC cancers. We observed that differences in clinical presentation and survival among cancer patients from different cohorts may be explained by bacterial abundance patterns. There are observed differences across racial groups. However, we are hesitant to conclude bacterial diversity or the relative abundance patterns impact racial disparities. In order to make such assertions racial minority groups' representation must be enriched in these of studies.
- We find that examining paired tumor and adjacent normal tissue is pivotal when evaluating the role of bacteria in the tumor microenvironment. Failing to do so can lead to wrongful interpretations. We note in our analyses that potential significant taxonomic associations disappear under strict paired conditions.
- Bacterial associations with cancer may be poly-microbial. Our study show the interaction between viral presence of HPV, HBV and HHV-4 and several bacterial species with overall survival as demonstrated in STAD, LIHC and HNSC cancers. This is supported by previous studies findings (Parsonnet 1995, Warren et al. 2013).
- Diversity relationship to overall survival is dependent on sex race and other clinical factors with opposite effects by per cancer type as demonstrated in STAD and LIHC. In STAD higher within sample bacterial diversity appears to have negative association with overall survival, while the association in LIHC is a positive one.

This study is not without its limitations, we found that bacterial reads were low in relation to human-host total number of reads which is an inherit restriction of working with human-derived whole exome sequencing data. We recognize that bacterial diversity and relative abundance divergent association with various tumor types, this does not signify causality. Our findings are limited to an association with the pathogenic process as determined by differential overall survival outcomes and further studies are needed to elucidate the mechanisms and the depth of the involvement. Also our pipeline design was unable to detect RNA viral reads which would have been important in the correlation analyses of viral interactions. We cannot rule out potential contamination. We examined submitter site and processing site and do not believe this to be a factor in our analysis. In addition, our strict paired design should control for potential contamination effect. We note that there is underrepresentation of racial minorities in our paired data subset of solid tumor cancers that hinders racial disparities analyses. Further, self-reported data and missingness including treatment information limit generalization of results. Nevertheless, validation of bioinformatics findings with the Hawaii Tumor Registry, an independent population derived from a diverse racial and ethnic population sample pool, strengthens our results.

4.2 **Future Directions**

As part of this project we downloaded and build microbial (viral and bacterial) profiles for 22 solid tumors from the Cancer Genome Atlas Network creating a platform for interdisciplinary collaboration that can

help move forward ongoing and future research. Presented here are the analyses for 9 of those cancers. These were selected on the basis on known infectious etiology, STAD, LIHC, HNSC, CESC and BLCA, or inflammation associated with no known infectious etiological factors, LUSC, LUAD, COAD and READ, that we considered key in identifying racial and ethnic differences in priority populations. Our efforts led to several abstract publications and presentations. This project will continue to analyze remaining 13 cancer cohorts and begin functional prediction analyses on the first 9 included here and continue to validate specific taxa in selected cancer cohorts. We can now better appreciate the diversity present in the human tumor microenvironment that makes identification of microbial communities so challenging. Nevertheless, this information can provide us with the necessary tools assess the effects of microbiota on the colonized tissue. Cross-cancer analyses such as this one, facilitate identification of inter-individual variation patterns and can answer additional questions about the beneficial or detrimental effects in different cancer lineages. Additional questions may include mechanistic approaches and mutational characterization, the prediction of clinically relevant functional differences and discovery of actionable pathways that can inform patient care and therapeutic approaches.

Possible Research Questions:

1. What are the clinical applications from microbial data derived from human sequencing reads and how can this help the understanding of racial variance and the impact on widening racial disparities?
2. Do bacterial relative abundance, diversity and co-occurrence patterns predict functional relevant inter-individual differences within the cancer type and across cancer types? If so, how can this best be applied in the application of primary and secondary prevention strategies and inform patient care
3. What are the mechanisms involved in microbe-host interaction effects and how can we best identify actionable pathways to reduce and eliminate racial and ethnic related disparities through microbial modulation?
4. Are bacteria diversity and relative abundance patterns associated with tumor somatic mutations and gene expression profiles? If so, what communities may be involved and to what degree? Are there any commonalities across cancer types or by organ systems?

CHAPTER V APPENDICES

Contents:

A. Supplemental Data tables
B. Data Management Plan & dbGaP Progress Report
C. IRB Approval Letter
D. R-package: main script and complete list of packages used
E. Literature cited (complete list)

A. **Supplemental data**

A.1 Core taxonomy across cancer types

Core Bacteria Species per cohort	Prevalence Tumor %	Prevalence Adjacent Normal %	Relative abundance Tumor % ±SD	Relative abundance Adjacent Normal % ±SD
STAD				
Bacillus subtilis	48	51	12.5	12.1
Mycoplasma mycoides	31	31	8.4	13.2
Cutibacterium acnes	31	29	16.3	19.6
Arthrobacter sp. IHBB 11108	22	29	16.8	6.4
Rothia mucilaginosa	26	24	7.8	6.2
LIHC				
Escherichia coli	67	67	89±30	84±31
Cutibacterium acnes	30	33	5±9	4±16
Ralstonia pickettii	31	32	3±21	4±30
COAD				
Escherichia coli	93	94	31±35	30 ±32
Bacteroides fragilis	81	86	14 ±20	10 ±14
B. thetaiotaomicron	73	81	6 ±10	6 ±7
Alistipes finegoldii	63	69	3 ±6	4 ±6
Bacteroides vulgatus	68	77	7 ±10	12 ±16
Parabacteroides distasonis	59	65	2± 4	3 ±6
Bacteroides dorei	63	75	4 ±9	5 ±9
Bacteroides ovatus	61	68	2 ±3	2 ±3
Bacteroides caecimuris	53	68	1 ±2	1 ±2
Roseburia hominis	60	60	1 ±2	1 ±1
Flavonifractor plautii	53	57	1 ±2	1 ±1
Cutibacterium acnes	53	55	1 ±2	1 ±4
READ				
Bacteroides dorei	31	36	29±15	32±12
Bacteroides vulgatus	44	72	25±7	37±19
B. thetaiotaomicron	67	67	38±28	44±29
Bacteroides fragilis	78	67	35±22	33±21
Escherichia coli	89	89	70±28	67±25
LUAD				
Bacillus subtilis	53	63	14±16	13±17
Mycoplasma mycoides	34	44	8±10	7±7
C. pseudotuberculosis	30	36	13±12	11±11
Arthrobacter sp. IHBB 11108	38	30	14±19	14±19
Cutibacterium acnes	25	22	18±18	18±18
Mitsuaria sp.7	26	29	12±10	19±20
R. depolymerans	23	22	8±9	14±8
Streptomyces gilvosporeus	24	28	5±4	5±5

Core taxonomy across cancer types (continued)

Core Bacteria Species per cohort	Prevalence Tumor %	Prevalence Adjacent Normal %	Relative abundance Tumor % ±SD	Relative abundance Adjacent Normal % ±SD
LUSC				
Bacillus subtilis	55	56	35±22	42±23
Mycoplasma mycoides	41	45	35±24	39±23
P. lacuslunae	37	36	34±23	42±23
C. pseudotuberculosis	33	37	35±22	40±24
Cutibacterium acnes	26	46	38±26	40±27
Streptomyces gilvosporeus	22	27	38±27	36±22
Bacillus mycoides	22	20	33±23	39±26
HNSC				
Bacillus subtilis	43	67	4±9	8±13
Mycoplasma mycoides	35	45	3±6	2±5
C. pseudotuberculosis	28	45	2±5	5+-9
Streptomyces gilvosporeus	22	33	1±2	2±4
Arthrobacter sp. IHBB 11108	23	46	1±4	5±9
CESC				
Escherichia coli	50	38	81±12	51±16
Bradyrhizobium BTAi1	38	25	46±7	91±9
S. koreensis	38	13	65±13	58±0
BLCA				
Bacillus subtilis	71	68	29±22	27±18
Arthrobacter sp. IHBB 11108	39	50	20±16	30±25
Mycoplasma mycoides	39	46	11±8	17±17
C. pseudotuberculosis	36	29	24±15	18±15

Percent of positive samples and average relative abundances (shown as percent) with plus/minus standard deviation in tumor and adjacent normal across nine TCGA cohorts. STAD: stomach adenocarcinoma; LIIC: liver hepatocellular carcinoma; COAD: colon adenocarcinoma; READ: rectal adenocarcinoma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; HNSC: head & neck squamous cell carcinoma; CESC: cervical squamous cell carcinoma; BLCA: bladder carcinoma.

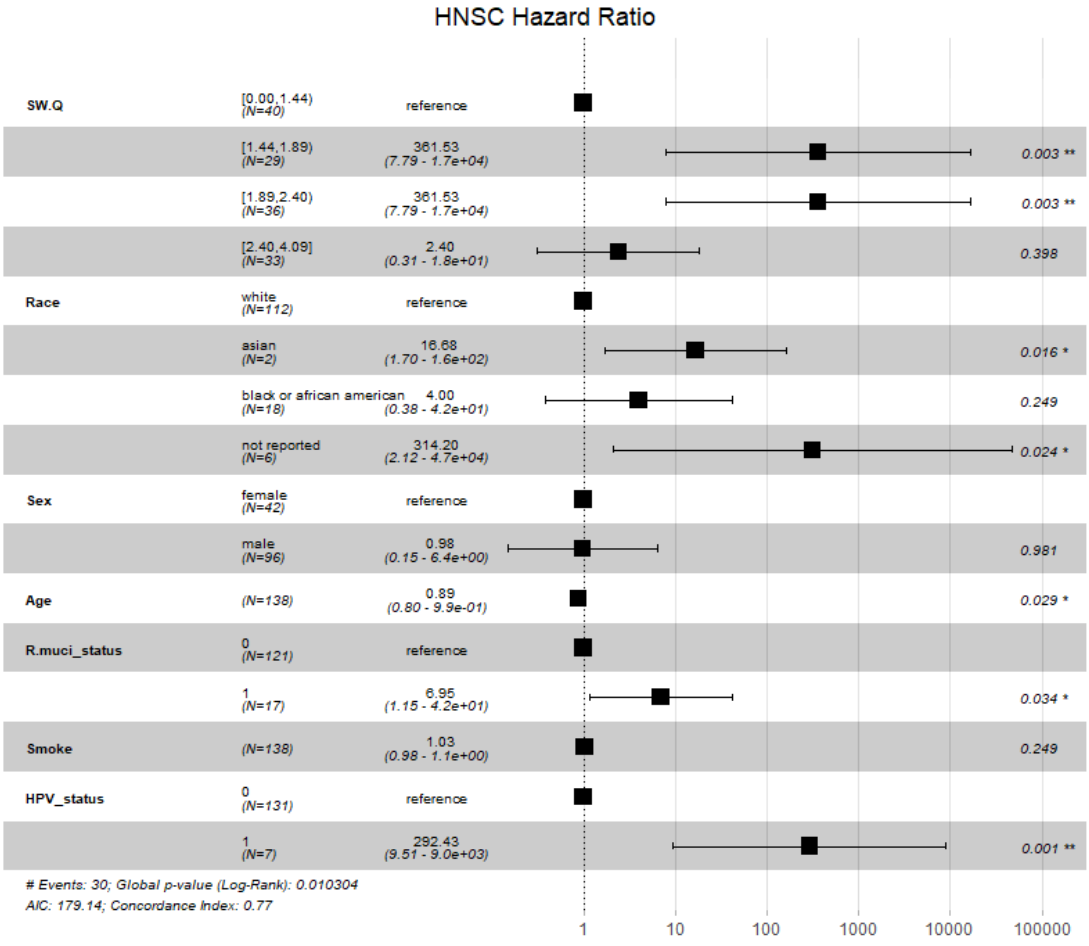
A.2 Diversity in cancers of the head & neck, liver and stomach

Variable	Reference	Estimate	SE	p-value	OR	95%CI
HNSC						
Intercept	Shannon Diversity	2.92	1.02	0.005	18.53	[2.45, 139.97]
Sample type	Adjacent normal	0.21	0.14	0.133	1.23	[0.94, 1.61]
Age at diagnosis		0.00	0.01	0.737	1.00	[0.98, 1.01]
Sex	Female	0.35	0.18	0.050	1.42	[1.00, 2.01]
Race	White					
Asian		-0.80	0.59	0.180	0.45	[0.14, 1.45]
Black or AA		0.15	0.24	0.529	1.16	[0.72, 1.88]
NR		0.79	0.43	0.066	2.21	[0.95, 5.14]
Ethnicity	Hispanic					
Not Hispanic		0.52	0.26	0.052	1.68	[0.10, 2.83]
NR		0.39	0.38	0.304	1.48	[0.70, 3.12]
Tumor Stage	stage i					
stage ii		-1.28	0.63	**0.047**	0.28	[0.08, 0.98]
stage iii		-0.67	0.64	0.291	0.51	[0.15, 1.79]
stage iv		-0.97	0.63	0.127	0.38	[0.11, 1.32]
Tumor grade	--	--	--	--	--	--
Anatomical site	Floor of mouth					
Hard palate		1.24	0.69	0.074	3.44	[0.88, 13.37]
Larynx		-0.93	0.38	**0.016**	0.40	[0.19, 0.84]
Lip,Oral,Pharynx, overlap		-0.60	0.40	0.135	0.55	[0.25, 1.21]
Tongue, base		-0.97	0.49	0.052	0.38	[0.14, 1.01]
Tngue, NOS		-0.92	0.39	**0.019**	0.40	[0.18, 0.86]
LIHC						
Intercept	Shannon Diversity	3.04	0.64	<0.0001	20.90	[5.94, 73.40]
Sample type	Adjacent normal	-0.17	0.17	0.328	0.85	[0.60, 1.18]
Age at diagnosis		-0.01	0.01	0.259	0.99	[0.98, 1.00]
Sex	Female	0.15	0.18	0.394	1.16	[0.82, 1.65]
Race	White					
Asian		-1.28	0.35	**<0.0001**	0.28	[0.14, 0.56]
Black or AA		-0.38	0.38	0.319	0.68	[0.32, 1.45]
NR		0.61	0.78	0.434	1.84	[0.40, 8.59]
Ethnicity	Hispanic					
Not Hispanic		-1.10	0.50	**0.028**	0.33	[0.13, 0.88]
NR		-1.89	0.84	**0.026**	0.15	[0.03, 0.79]
Tumor Stage	--	--	--	--	--	--
Tumor grade	--	--	--	--	--	--

Diversity (continued)

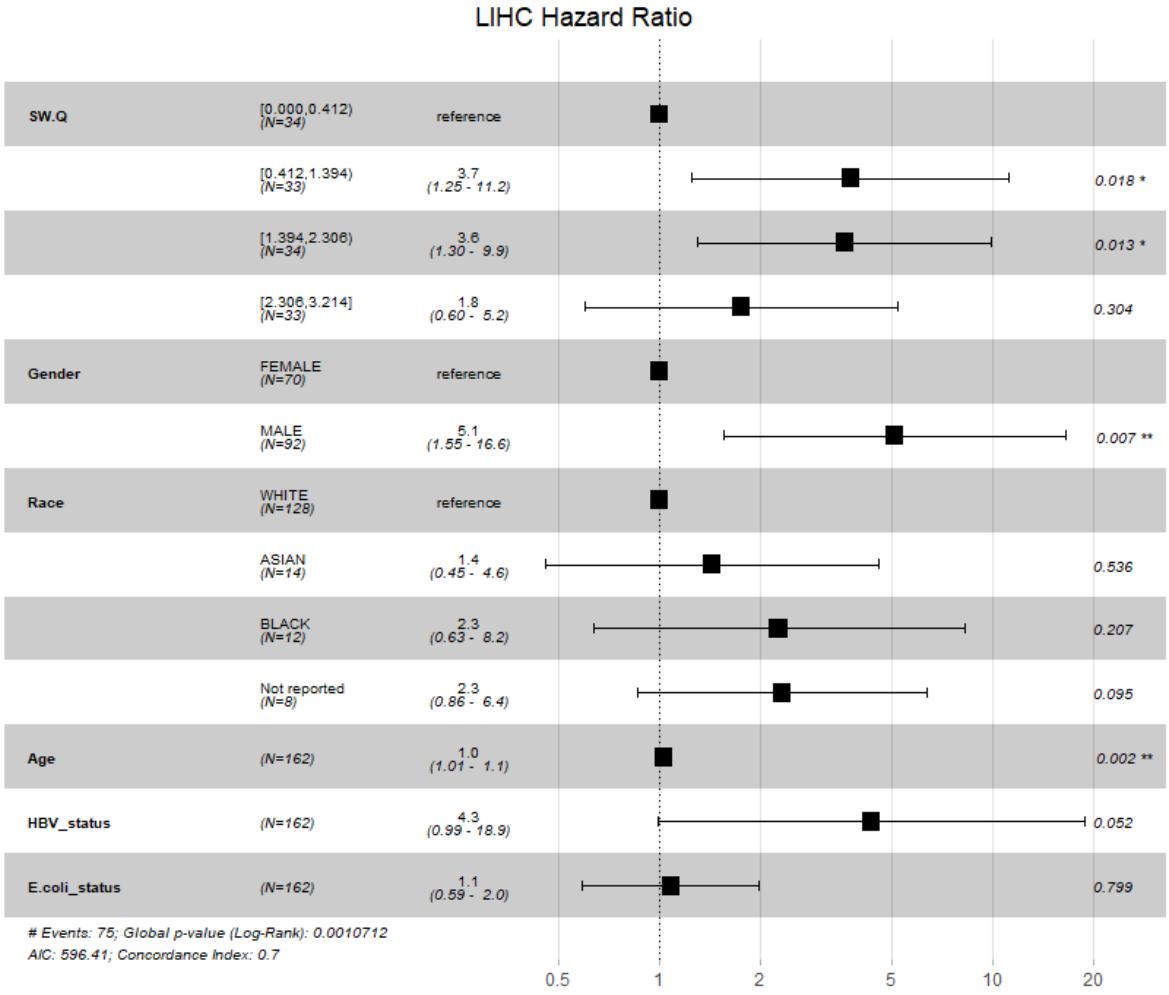
Variable	Reference	Estimate	SE	p-value	OR	95%CI
LIHC Anatomical site	--	--	--	--	--	--
STAD						
Intercept	Shannon Diversity	1.56	0.91	0.090	4.77	[0.78, 29.05]
Sample type	Adjacent normal	0.28	0.26	0.280	1.32	[0.79, 2.20]
Age at diagnosis		0.00	0.01	0.706	0.99	[1.00, 1.01]
Sex	Female	0.03	0.21	0.887	1.03	[0.68, 1.56]
Race	White					
Asian		-0.20	0.49	0.683	0.82	[0.31, 2.15]
Black or AA		1.26	0.59	**0.033**	3.53	[1.10, 11.29]
NR		-0.71	0.36	0.053	0.49	[0.24, 1.01]
Ethnicity	Hispanic					
Not Hispanic		NA	NA	NA	NA	NA
NR		NA	NA	NA	NA	NA
Tumor Stage	stage ia					
stage ib		0.27	0.49	0.587	1.30	[0.50, 3.42]
stage ii		0.84	0.55	0.131	2.31	[0.78, 6.86]
stage iia		0.42	0.46	0.369	1.52	[0.06, 3.81]
stage iib		0.57	0.05	0.231	1.78	[0.69, 4.57]
stage iii		1.93	0.96	**0.046**	6.87	[1.03, 45.80]
stage iiia		0.93	0.50	0.064	2.53	[0.95, 6.75]
stage iiib		1.19	0.57	**0.038**	3.29	[1.07, 10.10]
stage iiic		1.05	0.73	0.150	2.86	[0.67, 12.16]
stage iv		0.51	0.51	0.153	1.66	[0.60, 1.56]
Tumor grade	G1					
G3		0.33	0.30	0.273	1.39	[0.77, 2.51]
GX		0.05	0.70	0.460	1.67	[0.42, 6.64]
G norm		NA	NA	NA	NA	NA
Anatomical site	body of stomcah	NA	NA	NA	NA	NA
Cardia		-0.21	0.28	0.457	0.81	[0.47, 1.41]
Fundus		-0.58	0.82	-0.747	0.56	[0.11, 2.81]
gastric antrum		0.08	0.24	0.729	1.09	[0.67, 1.77]
stomach, nos		-0.63	0.49	0.201	0.54	[0.20, 1.40]

A.3 Hazards Ratio supplemental plots
HNSC



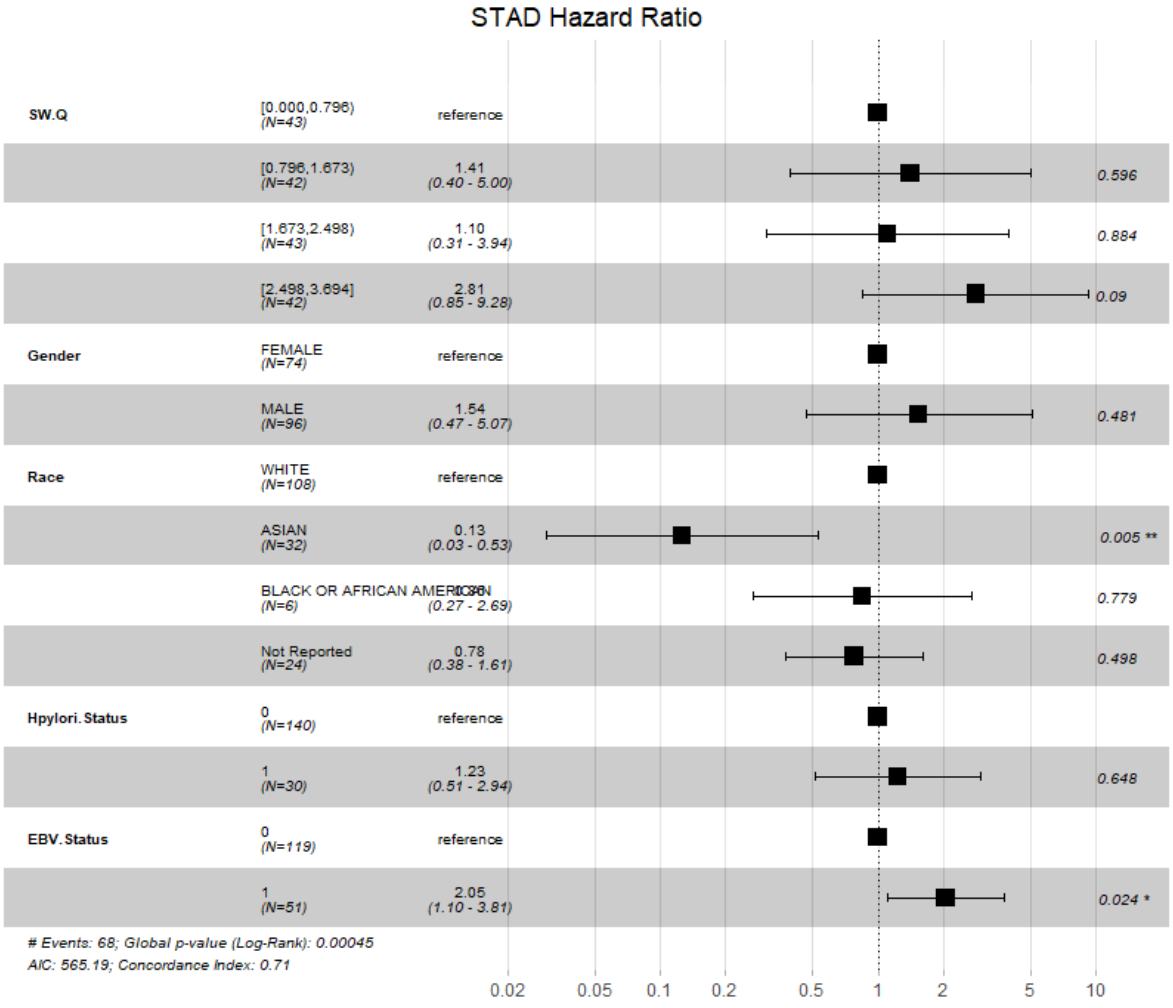
Cox proportional hazards forest plot illustrating effects of smoke in HNSC. When adding interaction of smoke, infection status of HPV and *Rothia mucilaginosa* have a detrimental effect on overall survival. However wide confidence intervals prevent accurate interpretation based on microbial presence or diversity levels in HNSC.

LIHC



Cox proportional hazards forest plot illustrating effects of smoke in LIHC. When controlling for interaction terms of age, race and sex significant hazards remain for males at intermediate diversity ranges.

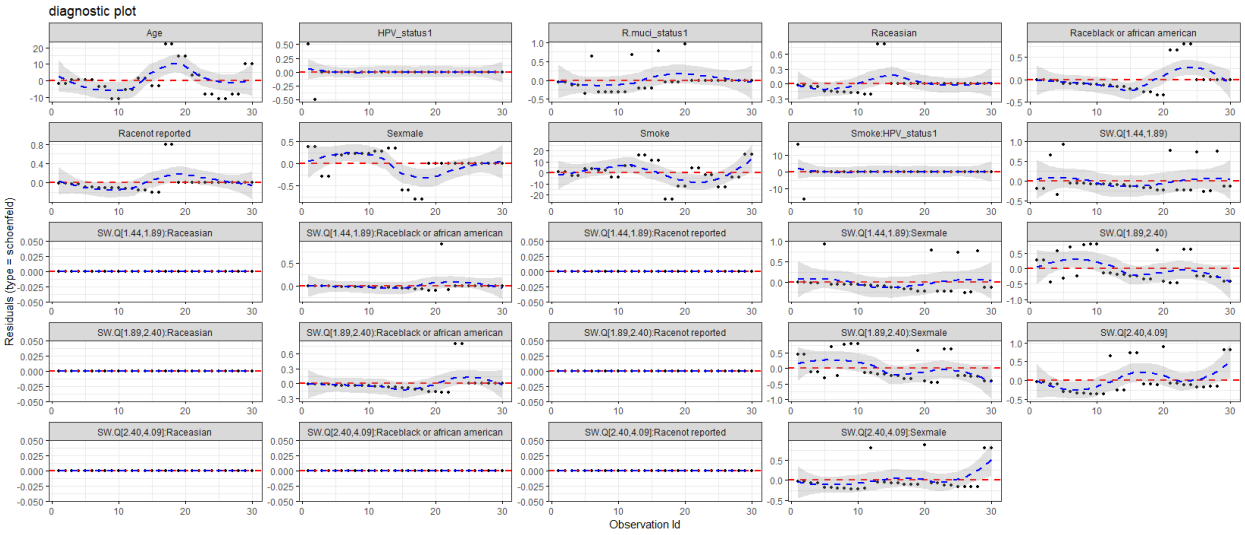
STAD



Cox proportional hazards forest plot for STAD controlling for interaction terms. After controlling for the interaction of sex and race, effects of bacterial diversity is no longer significant, while the effects of race and infection status remain. Among Asian patients who have on average lower diversity indices have significant lower risk compared to White counterparts (HR=0.13, 95%CI=0.03, 0.53, p=0.005).

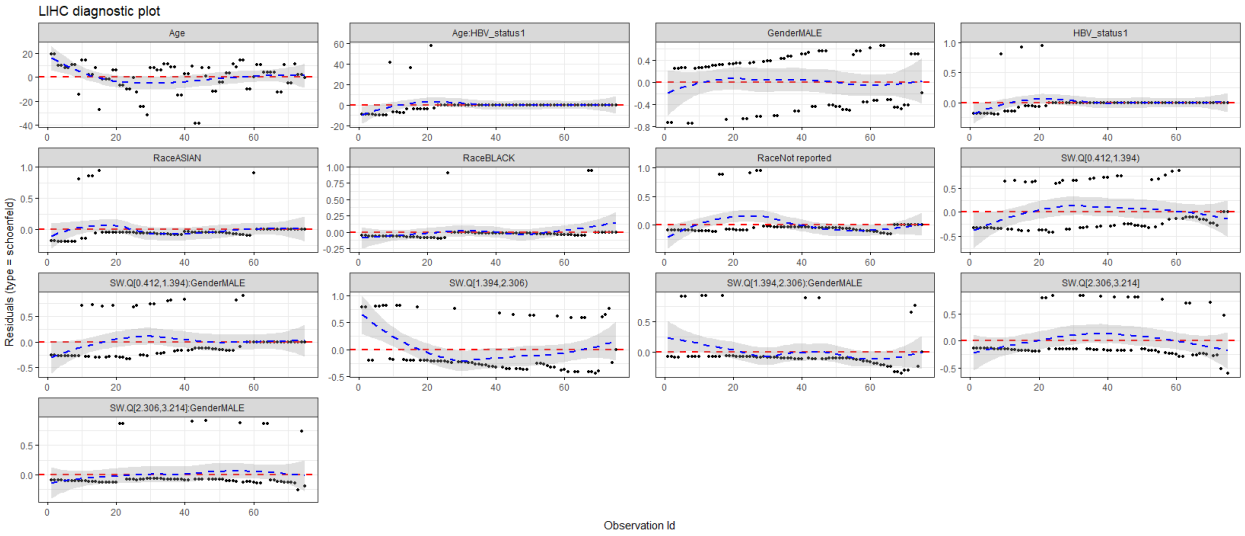
Diagnostic plots for Cox proportional hazards assumptions. Plots were created with survminer's ggcoxdiagnostic function. All plots include interaction terms.

HNSC diagnostic plots



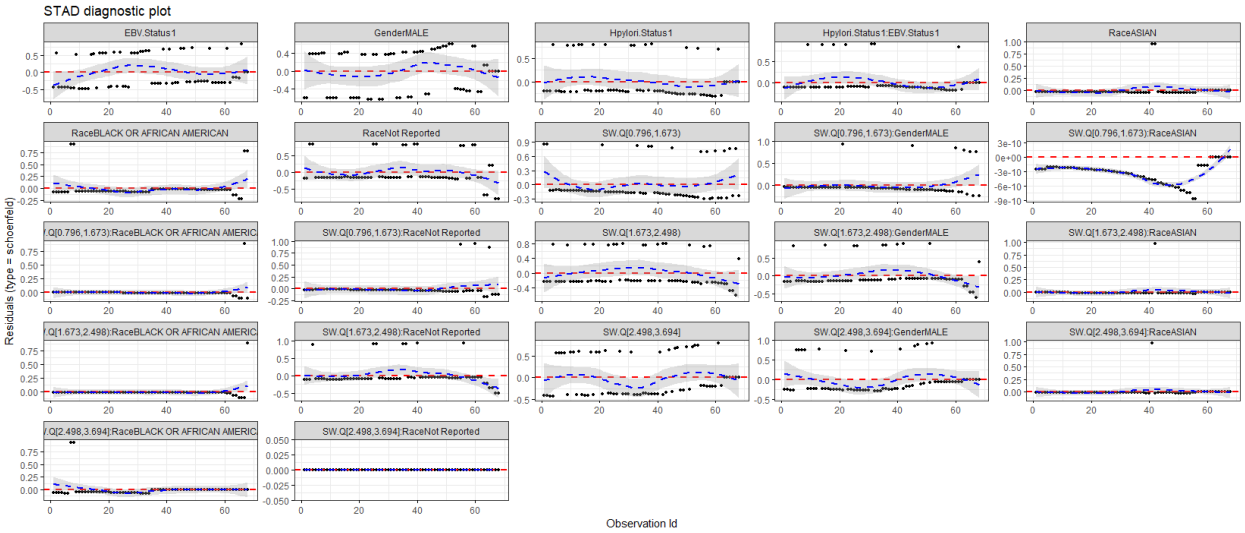
HNSC after correcting for proportional hazards violation by applying interaction terms of smoke, sex and race.

LIHC diagnostic plots



LIHC diagnostic plots after correcting proportional hazards violation by applying interaction terms of sex and infection status.

STAD diagnostic plots



Diagnostic plots for STAD after correcting for assumption violations with interaction terms of race, sex and infection status. Violation remains within Asian samples with intermediate-low diversity.

B. **Data Management**

B.1 Data Management Plan

This study was based on data generated from level 1 (original raw data) sample sequences from the TCGA consortium. As such a data management plan (DMP) was developed as part of the approval process. Original management plan is presented below. DMP has undergone 4 revisions to add additional researchers and additional datasets not related to this study.

TCGA Controlled-Access Data Management Plan

Protocol: Divergent Bacteria Associations with Cancer Pathogenesis across Tumor Types

DMP Version: 1.0

Date: 24 AUG 2017

Prepared by:

Youping Deng

Vedbar Khadka

Steve Mau

Mark Menor

Vance C Mizuba

Rebecca M Rodriguez

Reviewed by:

Leonard Gouveia, Institutional Signing Official

Vance C Mizuba, Institutional Information Technology Director

Overview:

This Data Management Plan (DMP) outlines policies and procedures related to data security as well as study database design and structure for our study (Project ID 14778) as it relates to the GDC Commons (dbGaP) data download and local storage of NIH Controlled-Access Data (phs000178.v9.p8).

This DMP is prepared by the Bioinformatics Core team in collaboration with the JABSOM IT Department. This DMP will be updated as necessary throughout the course of the study. Editorial changes and clarifications may be made without consultations. Once a change has been approved, the revised document will be made available with a new date and version number and implementation date if applicable. This plan will be reviewed on an annual basis. A copy will be maintained in the study regulatory folder.

Roles and responsibilities:

General protocol roles and responsibilities will apply. In brief, principal investigator (PI) is responsible for the procurement of NIH Controlled-Access Data used in this study, the protection of data confidentiality and of providing annual progress reports as delineated in the Data Use Certification (DUC). The Bioinformatics Core located at UH Manoa JABSOM Campus, is responsible for the management of the data. Institutional Signing Official (SO) and Information Technology (IT) Director are responsible for proper use and security of the data downloaded. As per dbGaP Security Best Practices, IT Director may not function as PI, SO or collaborating scientist. The following study key personnel and collaborators will have authorized access to the data in order to achieve the project's objectives:

Personnel	Project's objective
Youping Deng, Principal Investigator	*-Procurement of controlled access-data* *-Management of data integrity assurance* *-Preparation of process data for analysis, creation and execution of analytical programing* *-Annual Progress Report*
Brenda Y Hernandez, Study Collaborator	*-Procurement of controlled access-data* *-Validation of study findings* *-Annual Progress Report*
Mark Menor, Data Configuration and Programmer	*-Abstract raw data from NCI Genomic Data Commons* *-Preparation of process data for analysis, creation and execution of analytical programing* *-Management of data integrity assurance*
Vedbar Khadka Data Configuration and Management	*-Preparation of processed data for analysis, creation and execution of analytical programing* *-Management of raw/processed data integrity assurance*
Rebecca Rodriguez, Research Coordinator	*-Overall study coordination* *-Preparation of processed data for analysis, creation and execution of analytical programing*
Vance C Mizuba IT Director	*-Management of requested data set according to NIH expectations and Best Practices for Controlled-Access Data and UH Policies and Security Requirements* *-Develop and execute CAPA procedures in the event of data security incidents*

-Least Privileged Access. In order to access, process and analyze data the personnel listed above are granted access according to their roles to accomplish project's objectives. All personnel follows strict UH guidelines and policies to set unique user account and passwords.

-Personnel Training. All study staff are required to take appropriate training in information security awareness. The study coordinator ensures study personnel maintains compliance with training renewal as needed.

Definitions

This DMP is primarily for the download, processing, storage and security of TCGA human genomic Level 1 sequencing data. The data here referenced are whole genome DNA sequence binary alignment files (BAM), throughout this document described as "**raw data" or "controlled-access data**". These data are provided de-identified and will be used to create microbial profiles.

Clinical data, biospecimen data and pathology reports data are described in this document as "**clinical data**". Clinical data are being requested and will be downloaded from Level 1 (.xml) files. These data includes basic demographic information, treatment and survival data as well as details about of sample processing and pathology reports. These data are provided de-identified and will be used to correlate microbial profiles to clinical outcomes.

Human DNA sequences will be filtered out from raw data. Reads will be subsequently filtered, trimmed and compiled using validated pipeline as described (section 1). Remaining operational taxonomic units (mostly microbial DNA) will be aligned against reference genomes, throughout this document described as "**process(ed) data**"

1. Data Management Milestones

Due to the ongoing nature of this project, milestones relating to project completion and data lock are not included.

-**Pipeline Design.** Previously validated pipelines are being used with few modification to fit the needs of the project.

-**Pipeline Validation**. Pipeline specifications and programming were checked against open-access exome sequencing data for 9 gastrointestinal stromal tumors and normal matched pairs.

-**Controlled Access Data Download.** Request to controlled-data have been submitted via dbGaP (Project ID 14778). After data access approval notification, PI will grant access to key personnel according to their roles through dbGaP Portal. Download should commence within 5 business days upon granting access to key personnel.

-**Data Queries.** Data queries will be run once a month to ensure data integrity is maintained. Data configuration staff and the research coordinator will be responsible for resolving data queries.

-**Pilot Audit**. A pilot audit will be completed immediately following initial TCGA data download for the first 10 BAM files. This will ensure system is working as intended and files are downloading appropriately.

2. Data Collection Instruments

Raw data will be downloaded directly from NCI Genomics Data Commons in BAM files, processed and transcribed (processed data) into spreadsheets after the filtering process. This task will be completed by data configuration and/or programmers. Clinical data, as described in definitions sections, will be extracted from .xml files onto spreadsheets (when available). This task will be completed by the PI and/or research coordinator. Only project specific data will be retained as specified in the approved protocol, all other will be discarded as per UH Policies and Procedures HRS 487R. Raw data will be retained according to DUC Agreement for TCGA data and Institution data retention policies.

3. Collection Schedule

NIH Controlled-Access Data (TCGA, phs000178.v9.p8) and clinical data will be downloaded at once upon access approval and token authentication. PI will ensure key personnel is granted access to project data prior to requesting token authentication. Tokens are valid for thirty (30) days. It is estimated complete data download will take in excess of 30 business days and will depend on size of BAM files and speed of access. New tokens will be requested as needed per GDC Data User's Guide.

4. Data Flow

Data Source → Data Transfer → Data Management

1. NIH Controlled-access data will be downloaded from GDC Data Commons into Bioinformatics server and network storage system.
2. Data will be cleaned and transferred to study database for processing within the Bioinformatics servers.
3. Data will be stored and analyzed according to protocol to answer study aims.

5. Pipeline Design and Testing/Validation

The study pipeline and database will be setup and tested by the data configuration and programming team prior to approval for production use. All changes made to the pipeline structure/programming after final approval are to be documented in a change control log and included in an updated version of the project data management plan.

BAM to fastq → PathoQC → PathoMap → MegaBLAST filter → PathoID → PathoReport → taxize

6. Data Security

Protection of the privacy and confidentiality of the participants whom information is derived is our number one priority. Data security measures and appropriate safeguards will be employed to ensure compliance with standards as per UH policy, DUC Agreement and dbGaP Approved User Code of Conduct.

-Primary Work Locations (PWLs). The PWLs for this project are identified as:

UH Bioinformatics Core Biosciences Building, UH JABSOM Campus 651 Ilalo Street Honolulu, HI 96813	*MEB Server Room* Medical Education Bldg, UH JABSOM Campus 651 Ilalo Street Honolulu, HI 96813

Workstations and study staff reside at the UH Bioinformatics Core work location. The servers reside at the MEB Server Room location.

-**Physical Access Controls.** Access to the Biosciences building is restricted by key card access and manned 24/7 by security guards at the front desk. Access to the MEB server room is restricted by key card and biometric fingerprint access. Audit logs for ingress access are available as needed for both PWLs.

-**Environmental Protection.** The MEB server room has a FM-200 fire suppression system; UPS and generator power backup in cases of long-term power failures; and active monitoring for temperature range and water leakage.

-Data Storage. The server is the primary data storage location and houses the raw data and processed data sets. Study staff workstations may store copies of the processed data set as needed for daily workflows. All servers, workstations, and removable media that store data are encrypted-at-rest with full disk encryption enabled. The encryption uses 128-bit AES encryption cipher or better.

-Data Transfer. All data transfers over a network use end-to-end encryption methods, such as SSH or sFTP. All data transfer of processed data is made via password protected CD or encrypted transfer such as encrypted email or secure file transfer. Raw data stored on the server is prohibited from transfer.

Media Formats of Data		Safeguards
☒	Data stored on servers and workstations --Raw and processed data	Full disk encryption is implemented on the hard drive of all devices that will house NIH controlled-access data or a copy there of. Workstations are prohibited from storing raw data.
☒	Data on removable media (CD, DVD, portable hard drives, USB drives, etc.) --Processed data	All data on removable media storage will be encrypted with full disk encryption
☒	Data in printed format --Processed data for use in publications, presentations	Facility entrance is restricted. Physical precautions policies are in place to protect data in printed format to prevent unauthorized access.

-Encryption Key Management. All encryption keys for encrypted files and disks is stored either on a separate computer system not involved in this project or kept as hard copies in a locked secured file cabinet that only can be accessed by data management staff.

-Remote Access. Remote access away from PWLs is authorized through VPN secure connection. As per UH Policy, End-to-End encryption will be utilized on Secure Socket Layer (SSL) or Virtual Private Network (VPN) connections for transferring data. Only de-identified cleaned data (processed data) will be authorized remote access.

-Computer Network Technical Controls and Safeguards. The JABSOM Office of Information Technology (OIT) manages the network firewall, workstation firewalls, workstation group policy, and network intrusion prevention/detection system. Host based firewalls are required per UH policy on all workstations that store copies of the data.

Our computer Operating System (OS) are current with latest patches and security updates in accordance with the JABSOM OIT patch management policy. In brief, all Windows-based workstations are joined to the domain and receive security patches and updates automatically on a monthly basis. Per UH Policy, anti-virus/ anti-spyware software are deployed throughout the network on workstations and servers and are periodically updated.

Practices and policies are in place to safeguard servers and workstations during periods of inactivity. These include auto-logoff after 30 minutes of inactivity and screensaver lock after 15 minutes of inactivity for all workstations.

Security audit controls are implemented which record and examine user activities on the workstations where controlled-access data is processed. All logging attempts are tracked and auditable.

Vulnerability scans will be executed on a quarterly schedule. Remediation of all threats rated ‘high’ will be resolved in 7 days. Remediation of all threats rated ‘medium’ or ‘low’ will be resolved in 30 days.

In the event of a *security incident*, the PI and JABSOM OIT will be notified immediately. Steps to initiate corrective action or other remediation and mitigation to determine the data loss and prevent recurrence of the breach will be implemented immediately. The UH System Security Officer will also be notified.

-**Data Storage and Document Archiving**. Raw and processed data will be housed in a Network Attached Storage (NAS) server in a folder accessible only to the study team. No copies of the raw data will ever be made into any workstations. A study database will be compiled from the processed data (microbial and clinical data combined) for analyses. Study team will maintain copies of the study database on the bioinformatics core for at least one (1) year after analysis and according to UH record retention policies and according to DUC Agreement, section 6- *Data Security and Release Reporting*. After study completion and publication study files will be retained or discarded as per UH Policies and Procedures HRS 487R. Reports will be archived indefinitely as read-only files.

-Backup. Data backup will be conducted at least once a month at regular intervals by the data management study team. Backups are written to encrypted disks and maintained at PWL.

DMP Change Control Log

Date	Version	Author	Revision

References:

E2.210: Use and Management of Information Technology Resources Policy

E2.214: Security and Protection of Sensitive Information

E2.215: UH Institutional Data Governance Policy

HRS 487R: Destruction of Personal Information Records

NCI Genomic Data Commons (GDC) Data User's Guide

NCI Genomic Data Commons (GDC) Data Transfer Tool User's Guide

The Cancer Genome Atlas (TCGA) Data Use Certification Agreement, version 20.AUG.2014

NIH Genome Wide Association Studies (GWAS) Data Sharing Policy

dbGaP Security Best Practices

B.2 dbGaP Data Access Request (DAR)

Project Renewal

Project #14778 : Divergent Bacteria Associations with Cancer Pathogenesis across Tumor Types

Project name	Divergent Bacteria Associations with Cancer Pathogenesis across Tumor Types
Project ID	14778
Approved user name	Youping Deng
Institute affiliation	UNIVERSITY OF HAWAII AT MANOA (Non-Profit)
Request ID	57292-4
Request date : 2018-10-09	Renewal date :

Applicant Organization

Legal Name :	**UNIVERSITY OF HAWAII AT MANOA**						
Department :	Complimentary and Integrative Medicine			Division :	Bioinformatics		
Street 1 :	651 Ilalo Street						
City :	Honolulu	State :	HI	Zip :		Country :	United States

PI Contact Information

Name :	**Youping Deng**			Position :	Principal Investigator		
Organization:	UNIVERSITY OF HAWAII AT MANOA						
Street 1 :	University of Hawaii John A. Burns School of Medic				651 Ilalo Street		
City :	Honolulu	State :	Hawaii	Zip :	96813	Country :	USA
Phone :	8089621664	:		Email :	dengy@hawaii.edu		

SO Contact Information

Name :	**James Ash**			Position :	Signing Official		
Organization:	UNIVERSITY OF HAWAII AT MANOA						
Street 1 :	Office of VP for IT and CIO				2520 Correa Rd, IT Center, 6th Floor		
City :	Honolulu	State :	HI	Zip :	96822	Country :	United States
Phone :	808-956-7241	:		Email :	jtash@hawaii.edu		

IT Director Contact Information

Name :	**Vance Mizuba**			Position :	Director of Information Technology		
Organization:	UNIVERSITY OF HAWAII AT MANOA						
Department :	John A Burns School of Medicine			Division :	OIT		
Street 1 :	651 Ilalo St						
City :	Honolulu	State :	Hawaii	Zip :	96813	Country :	United States
Phone :	8086921116	:		Email :	vancecm@hawaii.edu		

Project Renewal

Project #14778 : Divergent Bacteria Associations with Cancer Pathogenesis across Tumor Types

Approved Research Use Statement

Infections are thought to account for up to 20% of the total cancer burden worldwide. Contribution to the cancer burden by bacterial pathogens is underestimated. To better understand potential bacterial associations with different cancer types, we propose an integrated data analysis followed by experimental validation with at least 8,000 tumor-normal matched read-pairs across all tumor types from TCGA raw sequencing data (BAM files). To this end we request access to dbGaP, TCGA consortium data. Unaligned reads will be extracted using validated methods and filtered for human content. Non-human read-pairs will be aligned to reference bacterial genomes and relative abundances estimated to build profiles. We plan to correlate microbial differences across tumor types with clinical data to determine bacterial associations with cancer modulation effects that can be explored for potential tumor-dependent therapies. We will also perform follow-up experimental validation analyses to confirm findings. Our aims are consistent with the General Research Use governing the use of the data selected and TCGA efforts. We will adhere to outlined policies and protection of the privacy and confidentiality of participants whom data were collected from. We do not anticipate this study will pose additional risks.

Non-Technical Summary

The overall goal of this project is to correlate differences in microbial composition across tumor types to clinical outcomes. Comparison of the microbial profiles across tumor types and their correlation to clinical outcomes will enable further elaboration on their carcinogenic role by cancer type. Long term goal is to help determine bacterial contributions to cancer modulation and potential tumor specific target therapies.

Collaborators

Internal

Brenda Hernandez
Associate Researcher and Full Member
UNIVERSITY OF HAWAII AT MANOA -- Department: Cancer Epidemiology -- Division: UH Cancer Center
701 Ilalo St
Honolulu, Hawaii 96813 United States
Phone: 8085862992 Email: brenda@cc.hawaii.edu

Change Log

Date	Changed Details
2018-10-09 08:00	Presentations
2018-10-09 07:00	Research Progress
2018-10-09 06:00	Research Progress
2018-10-09 06:00	Presentations
2018-10-09 06:00	Inappropriate Data Use
2017-09-14 04:00	Signing Official
2017-08-10 18:00	Collaborators
2017-08-10 17:00	Research Use Statement
2017-07-14 05:00	Title
2017-07-14 04:00	Research Use Statement, Public Summary
2017-07-14 01:00	Collaborators
2017-07-14 01:00	Research Use Statement, Public Summary

Research Progress

Research Summary

Project Renewal

Project #14778 : Divergent Bacteria Associations with Cancer Pathogenesis across Tumor Types

This project seeks to correlate microbial relative abundance to clinical features and survival data by comparing whole exome sequencing data between TCGA tumor and adjacent paired tissue samples to understand if bacteria play an active role in various cancer types. We downloaded DNA exome sequences from the TCGA data set phs000178.v10.p8 as per data user agreement and data management plan. Thus far, Eighty-eight (88) stomach adenocarcinoma (STAD) and eight(8) cervical squamous cell carcinoma & endocervical adenocarcinoma (CESC) TCGA cohort cases have been evaluated. TCGA genomic sequencing data was processed through a bioinformatics pipeline designed to generate microbial profiles from DNA sequences (BAM file format) using PathoScope 2.0. Clinical features were derived from TCGA specimen metadata files. Diversity indices for these cancers were calculated to compare microbial taxa diversity within tumor samples and within paired normal adjacent samples using R software. We have found differences in microbial diversity among the racial groups in tumor and adjacent normal samples as well as common taxa for both of the cancers evaluated (B. subtilis and C. acnes) indicative of core taxa or potential contamination and is being further evaluated. In addition we have found that some microbial species known to be over represented in cancer compared to non-cancer such as H. pylori are over represented in adjacent tissue compared to tumor in agreement with previous findings by Tang et al. 2005 (PMC4250691) demonstrating the importance of evaluating adjacent tissue samples. We have found that microbial information derived from exome sequencing are comparable to 16S and RNA-seq methods we have been unable to complete survival analyses due to sparse data from selected paired samples. Since our initial approval we have had one oral poster presentation and no other accepted publications.

Scientific Presentations

Poster presentation: Bacterial Differential Abundance Derived from Tumor and Solid Normal Tissue DNA Sequences May Help Explain Health Disparities in Certain Cancers. Rodriguez R, Khadka V, Menor M, Deng Y, Hernandez B. Biomedical Sciences and Health Disparities Symposium John A Burns School of Medicine University of Hawaii Cancer Center. April 18-18, 2018.

Publications

none

Intellectual Property

none

Data Security

Datasets not described in the Research Use Statement

none

Inappropriate Data Use

None

Project Renewal

Project #14778 : Divergent Bacteria Associations with Cancer Pathogenesis across Tumor Types

Consent Group Information

phs000178.v10.p8 : The Cancer Genome Atlas (TCGA)

DAR : 57292

Consent Group : 1

Name : General Research Use
Abbreviation : GRU
Request Date : 2017-07-10 Last Renewal Date :
Use Limitation : Use of the data is limited only by the terms of the model Data Use Certification.

C. **IRB**

C.1 IRB approval letter

UNIVERSITY of HAWAI'I SYSTEM

Office of Research Compliance
Human Studies Program

TO:	Khadka, Vedbar, PhD, Complementary and Alternative Medicine, University of Hawaii at Manoa Okiyama, Eugene, MPH, University of Hawaii at Manoa, University of Hawaii Cancer Center, Hernandez, Brenda, PhD, MPH, University of Hawaii at Manoa, University of Hawaii Cancer Center, Rodriguez, Rebecca, MS, PBT (ASCP) , CCRC, Complementary and Alternative Medicine, Area Health Education Center, University of Hawaii at Manoa
FROM:	Rivera, Victoria, Interim Dir, Ofc of Rsch Compliance, Biomedical IRB
PROTOCOL TITLE:	Bacterial Association with Cancer Pathogenesis Using TCGA Sequencing Data
FUNDING SOURCE:	NIH/NIMHD
PROTOCOL NUMBER:	2018-00174
	Approval Date: April 16, 2018 Expiration Date: December 31, 2999

NOTICE OF APPROVAL FOR HUMAN RESEARCH

This letter is your record of the Human Studies Program approval of this study as exempt.

On April 16, 2018, the University of Hawaii (UH) Human Studies Program approved this study as exempt from federal regulations pertaining to the protection of human research participants. The authority for the exemption applicable to your study is documented in the Code of Federal Regulations at 45 CFR 46.101(b) 4.

Exempt studies are subject to the ethical principles articulated in The Belmont Report, found at the OHRP Website www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.

Exempt studies do not require regular continuing review by the Human Studies Program. However, if you propose to modify your study, you must receive approval from the Human Studies Program prior to implementing any changes. You can submit your proposed changes via email at uhirb@hawaii.edu. (The subject line should read: Exempt Study Modification.) The Human Studies Program may review the exempt status at that time and request an application for approval as non-exempt research.

In order to protect the confidentiality of research participants, we encourage you to destroy private information which can be linked to the identities of individuals as soon as it is reasonable to do so. Signed consent forms, as applicable to your study, should be maintained for at least the duration of your project.

This approval does not expire. However, please notify the Human Studies Program when your study is complete. Upon notification, we will close our files pertaining to your study.

If you have any questions relating to the protection of human research participants, please contact the Human Studies Program by phone at 956-5007 or email uhirb@hawaii.edu. We wish you success in carrying out your research project.

1960 East-West Road
Biomedical Sciences Building B104
Honolulu, Hawai'i 96822
Telephone: (808) 956-5007
Fax: (808) 956-8683
An Equal Opportunity/Affirmative Action Institution

D. **R-Package tools**

D.1 R script (Microbial differential abundance-main)

```
##############################################################################################
#'script for TCGA microbiome differential analyses starting with phyloseq object#####
##############################################################################################
#load libraries
library(phyloseq)
library(ape)
library(microbiome)
library(ggplot2)
library(vegan)
##############################################################################################
#Part I. Build phyloseq object#################################################################
##############################################################################################
#A. Load and arrange data needed to build phyloseq object
#1. otu_table and taxa_table
##IMPORTANT: before loading the file with the otu counts we should have cleaned the file##
##meaning:
##should not contain any records of patients without bacterial reads. Leaving them will##
##change the results as it will erroneously have all bacteria present in as significant##
##number of records##

otucount <- read.csv('XXXX_bacteria_counts.csv', row.names = 1)+1 #before running
#script ensure to change "XXXX for the name of the cancer type i.e STAD"
#Note: addition of 1 is important to avoid errors with log of zero
otutable<-data.matrix(otucount, rownames.force = NA)
head(otutable)[1:5]
class(otutable)

#remove microbial reads that are not present in both tumor and normal across all#
#samples…should not be different if adding 1. If running for creation of relative#
#abundances, DO NOT add 1#######
raw <- rowSums(otucount)
otucountC <- otucount[raw != 0,]
dim(otucountC)

taxdata<- read.csv("XXXX_taxa_table.csv",sep=',', header = T, row.names = 1)#before
runing script ensure to change "XXXX for the name of the cancer type i.e STAD"
taxtable <- as.matrix(taxdata, rownames.force = row.names(otutable))
head(taxtable)
class(taxtable)

#2. sample metadata
sampledata<- read.csv("XXXX_sample_meta.csv", sep=',', header= T, row.names= 1)
#before runing script ensure to change "XXXX for the name of the cancer type i.e #STAD"
sampledata<- sample_data(data.frame(sampledata))
```

```
#3. merge initial phyloseq object (otu and taxa) to create rooted tree
OTU <- otu_table(otucountC, taxa_are_rows = TRUE)
TAX <- tax_table(taxtable)
head(OTU)
head(TAX)

XXXXphyloseq <- phyloseq(OTU, TAX) #before runing script ensure to change "XXXX for
the name of the cancer type i.e STAD"
XXXXphyloseq #before runing script ensure to change "XXXX for the name of the cancer
type i.e STAD"


#4. rooted tree
XXXXtree            <-            rtree(ntaxa(XXXXphyloseq),            rooted=TRUE,
tip.label=taxa_names(XXXXphyloseq), br= runif(1))#before runing script ensure to
change "XXXX for the name of the cancer type i.e STAD"
plot(XXXXtree)
XXXXtree

#B. Build phyloseq object summarized experiemnt with all components
XXXXpseq <- merge_phyloseq(XXXXphyloseq, sampledata, XXXXtree)#before runing script
ensure to change "XXXX for the name of the cancer type i.e STAD"
XXXXpseq
################################################################################
#Part II. Use phyloseq object to obtain microbial densities with microbiome R-package
################################################################################
#####load any mising dependencies#####

#A. create relative abundance table from pseq object and save file for future analyses
#remember to change XXXX for the cancer type before running the script#
XXXXpseq.compositional <- transform(XXXXpseq, "compositional")
XXXX_relabund<-XXXXpseq.compositional@otu_table
write.csv(XXXX_relabund, file = "XXXX_ra_from_pseq_object.csv")#double check no
samples get dropped.
#microbiome package will drop to make table with even number of samples.

#B. Identify core taxa. May have to change the detection and prevalence levels until#
#dataframe produces values.
####OTU abundance data must have non-zero dimensions.##############################
################################################################################
#if receiving above error message change the prevalence reducing 10%ges at the time#
#until error dissapears#########################################################
#Core taxa ideally is Taxa with over 50% prevalence at .2% or .5 % relative #abundance.#
#Viral study (Cantalupo et al. Virology. 2018; 513:208-216) included any and all virus#
#in at least 5% of the population without defining true core. Our internal process#
#includes any taxa with 20% prevalence in either tissue type as determined by read
#counts with a relative abundance of min# #.2% in at least 50% of those positive for#
#the microbe therefore there is room to play around to determine core taxa for each#
#cancer type in #order to compare across#########################################
```

```
#############################################################################################
#1. core taxa by relative abundance (compositional transformation)
XXXXcore <- core(XXXXpseq.compositional, detection = .2/100, prevalence =
20/100)#before runing script ensure to change "XXXX for the name of the cancer type
i.e STAD"
XXXXcore

#before running script ensure to change "XXXX for the name of the cancer type i.e STAD"
XXXX_p <- plot_core(transform(XXXXcore, "compositional"),
               plot.type = "heatmap",
               colours = gray(seq(0,1,length=5)),
               prevalences = seq(.05, 1, .05),
               detections = 10^seq(log10(1e-3), log10(.2), length = 10),
               horizontal = TRUE) +
  xlab("Detection Threshold (Relative Abundance (%))")
print(XXXX_p)

#2. collect the tids of the taxa resulted from the core plot
#3. plot core taxa relative abundance in the population
#before running script ensure to change "XXXX for the name of the cancer type i.e STAD"
plot_composition(transform(XXXXcore, "compositional"),
                 plot.type = "barplot", sample.sort = "neatmap")

#C. Ordinate data using the phyloseq object and microbiome R-package
#remember to set the seed. we are using the seed "4235421" and the Bray method per
#microbiome package tutorial

#1. set seed
set.seed(4235421)
ord <- ordinate(XXXXpseq, "MDS", "bray")#remember to change XXXX to the cancer type
before running the script

#2. by Sample type
plot_ordination(XXXXpseq, ord, color = "Sample.Type") +
  geom_point(size = 5)

#3. may copy above sentence code with different variables as needed

#4. canonnical (CCA) ordination may be more informative when looking a % explained by
racial differences for example or taxa variations
#4a. by race
XXXXpseq.cca <- ordinate(XXXXpseq, "CCA")
p <- plot_ordination(XXXXpseq, XXXXpseq.cca,
                     type = "Sample.Type", color = "Race")
p <- p + geom_point(size = 4)
print(p)

#4b. by taxonomy
p <- plot_ordination(XXXXpseq, XXXXpseq.cca,
```

```
                    type = "taxa", color = "phylum")
p <- p + geom_point(size = 4)
print(p)

#5. if interested may plot density of the core taxa or of the significant taxa
individually
#using the tid or the name according to how it was recorded in the otu table
#5a for absolute (observed abundance)
plot_density(XXXXpseq, "XXXXxx") +ggtitle("Absolute abundance of 'enter name of taxa'
")
##remeber to change XXXX for the cancer type, XXXXxx for the tid and to enter the name
of the taxa in the title
##before running the script

#5b for relative abundances
x <- microbiome::transform(XXXXpseq, "compositional")
tax <- "XXXXxx"
plot_density(x, tax, log10 = TRUE) +
  ggtitle("Relative abundance 'enter name of taxon' ") +
  xlab("Relative abundance (%)")
##############################################################################
#Part III. Diversity Metrics using vegan package
##############################################################################
#A. load data in from counts table with bacteria counts only
##table must be flipped with taxa in columns (taxa=columns) may use the transpose
function t()
##it may also be necessary to load a separate sample metadata or annotation file for
plotting

#1. load data
XXXX_meta<-read.csv("XXXX_sample_meta.csv")
XXXX_bac.counts<-read.csv("XXXX_bacteria_counts_flip.csv",  row.names  =  1)#analyses
#require data be supplied as taxa=columns
str(XXXX_bac.counts)# can use to check if ny are not integers that may need adjusting
if so may use ceiling function

XXXX_bac.counts<-ceiling(XXXX_bac.counts) #remember to change the XXXX for the cancer
type

XXXX_bac.anno<-read.csv("XXXX_bacteria_annotation.csv")#may   use   sample   meta   or
specific bacteria annotation if sample size is different

#2. bind by column the counts and the annotation
XXXX_data<-cbind.data.frame(XXXX_bac.counts, XXXX_bac.anno)#ensure samples were sorted
correctly

#3. Calculate Diversity indeces, Shannon, Simpson, invSimp etc. and get the stast
##remember to change XXXX for the cancer type before running the script
XXXX.Simpson<-diversity(XXXX_bac.counts, index = "simpson")
```

```
summary(XXXX.Simpson)#provides the statistics for the samples

XXXX.Shannon<-diversity(XXXX_bac.counts, index = "shannon")
summary(XXXX.Shannon)#provides the statistics for the samples

XXXX.invSimp<-diversity(XXXX_bac.counts, "inv")
summary(XXXX.invSimp)


#4. Calculate the fisher alpha, should be invariant to sample size
XXXX.Fisher<-fisher.alpha(XXXX_bac.counts)
XXXX.Fisher

#5. Run a pairwise with all diversity metrics including fisher alpha
pairs(cbind(XXXX.Shannon,XXXX.Simpson,  XXXX.invSimp,  XXXX.unbSimp,  XXXX.Fisher),
pch="o", col="dodgerblue")#only include Fisher if non-NANs produced

#6. Determine speies counts by number of reads and richness (S)
XXXX.sr<-rowSums(XXXX_bac.counts)#total short reads per sample that are assigned to
individual taxa output is species count
#may save output as a csv to add column in the metadata or print to screen with quotes
to copy and paste

reads<-as.data.frame(XXXX.sr)
print(reads, quote = TRUE)# copy and paste output to excel file or add to metadata

#7. Determine richness and evenness save file for metadata addition
XXXX.S<-specnumber(XXXX_bac.counts)#richness "S" where S is the number of species

XXXX.eH<-XXXX.Shannon/log(XXXX.S)# evenness defined by Shannon's index (H)
#it may be important to compare simpson's and shannon's measures

#8. Calculate Beta diversity as defined by gamma/alpha-1
#remember to change XXXX to the cancer type before running script
XXXX.alpha<-with(XXXX_bac.anno,   tapply(specnumber(XXXX_bac.counts),   Sample.Type,
mean))
XXXX.gamma<-with(XXXX_bac.anno, specnumber(XXXX_bac.counts, Sample.Type))
XXXXBeta<-XXXX.gamma/XXXX.alpha-1
XXXXBeta #provides mean beta diversity per sample type, for indiv. change samp.type
#for sample ID
XXXX.alpha
XXXX.gamma #compare gamma with data from pivots, should be same composition#

#print the beta diversity by primary tumor and solid normal
#9. plot histograms and boxplots of the diversity indices to observe the breakdown
hist(XXXX.Simpson, col="dodgerblue", main = "XXXX.Simpson" )
hist(XXXX.Shannon,col = "orange", main= "XXXX.Shannon")

#observe div indices, if warranted, then create ggplot
```

```
par(mfrow=c(1,5))
boxplot(XXXX.Shannon~ XXXX_bac.anno$Sample.Type, main="XXXX Shannon Diversity Index",
xlab="Sample Type", ylab="Diversity Index")
boxplot(XXXX.Simpson~ XXXX_bac.anno$Sample.Type, main="XXXX Simpson Diversity Index",
xlab="Sample Type", ylab="Diversity Index")
boxplot(XXXX.invSimp~ XXXX_bac.anno$Sample.Type, main="XXXX invSimp Diversity Index",
xlab="Sample Type", ylab="Diversity Index")
boxplot(XXXX.unbSimp~ XXXX_bac.anno$Sample.Type, main="XXXX unbSimp Diversity Index",
xlab="Sample Type", ylab="Diversity Index")
boxplot(XXXX.Fisher~   XXXX_bac.anno$Sample.Type,   main="XXXX   Fisher's   alpha",
xlab="Sample Type", ylab="Diversity Index")
#it may be that only simpson and shannon are needed

#10. save files of diversity indices to include with metadata for future analyses

XXXX_diversity_indeces<-cbind(XXXX.Shannon, XXXX.Simpson, XXXX.invSimp, XXXX.unbSimp,
XXXX.Fisher, XXXX.sr, XXXX.S, XXXX.eH)
write.csv(XXXX_diversity_indeces, file= "XXXX_diversity_indeces.csv")

#Part IV Differential abundance by wilcox() test
#testing will be completed for each tissue separately and then combined. It is IMPORTANT
to ensure data is sorted#
#when loading data for each tissue type ensure only samples with bacterial data are
included and that the aggregate at species #
#level data is being loaded#
#remeber to change XXXX for the cancer type before running script

#load data
tumorRA<-read.csv("XXXX_tumor_bacteria_relabund.csv", row.names = 1)
head(tumorRA)
normalRA<-read.csv("XXXX_normal_bacteria_relabund.csv", row.names = 1)
head(normalRA)

#A. Tumor relative abundance testing
#1. mean and sd measures
T_mean_RA<-apply(tumorRA, 1, mean)
T_sdRA<-apply(tumorRA, 1, sd)
measure<-cbind(T_mean_RA, T_sdRA)

write.csv(measure, file = "XXXX_tumor_bacteria_aggregate_mean_sd.csv")

#plot histogram with log 10 of abundace to observe distribution
hist(log10(apply(tumorRA, 1, var)),
     xlab="log10(variance)", breaks=50,
     main="Log Transformed OTUs Relative Abundance Variance- Tumor Bacteria")

#2. wilcoxon one sample test
colnames(tumorRA)#verify column names
idx_1 <- grepl('.*\\.01', colnames(tumorRA))
```

```
pval<-tumorRA$pval <- apply(tumorRA, 1, function(x) wilcox.test(as.numeric(x[idx_1]),
paired = F, exact=F)$p.value)
#paired test should be set to false for one sample test, exact=F to handle zeros

padj<-p.adjust(pval, method = "fdr")#we are using fdr correction although it may be
unecessary due to small sample size of our data sets
head(pval)
head(padj)

tumorBACsigtab<-cbind(tumorRA, padj)
head(tumorBACsigtab)
write.csv(tumorBACsigtab,                              file                              =
"Results_XXXX_tumor_bac_only_RA_wilcox_sigtab_pval_padj.csv")#remember to change XXXX
#for cancer type

#B. Normal relative abundance testing
#1. mean and sd measures
N_mean_RA<-apply(normalRA, 1, mean)
N_sdRA<-apply(normalRA, 1, sd)
measure2<-cbind(N_mean_RA, N_sdRA)

write.csv(measure2, file = "XXXX_normal_bacteria_aggregate_mean_sd.csv")

#plot histogram of normal tissue
hist(log10(apply(normalRA, 1, var)),
     xlab="log10(variance)", breaks=50,
     main="Log Transformed OTUs Relative Abundance Variance- Adjacent Normal Bacteria")

#2. wilcoxon one sample test
colnames(normalRA)#verify column names
idx_11 <- grepl('.*\\.11', colnames(normalRA)) ## ensure all

pval<-normalRA$pval         <-         apply(normalRA,         1,         function(x)
wilcox.test(as.numeric(x[idx_11]), paired = F, exact=F)$p.value)
padj<-p.adjust(pval, method = "fdr")
head(pval)
head(padj)

normalBACsigtab<-cbind(normalRA, padj)

write.csv(normalBACsigtab,                              file                              =
"XXXX_normal_bac_only_RA_wilcox_sigtab_pval_padj.csv")#remember  to  change  XXXX  for
cancer type

#C. Combined differential abundance testing using wilcoxon paired test
#1. load data; ensure data is sorted! dataset should include all samples even empty
columns, meaning all paired
```

```
#samples with at least one microbe in one tissue type should include both pairs. if
#both pairs have no microbes (tumor and normal) DO NOT INCLUDE!#

XXXX_RA_data <- read.csv('XXXX_overall_bacteria_relabund.csv', header= TRUE,row.names
= 1) + 0.0001 #arbitrary number to manage log of zero, change according to data set
#to the lowest abundance present within the file. Addition of arbitrary number will
#affect lof2fc incrementally. Run with and without compare true values.
XXXX_RA_data <- data.frame(XXXX_RA_data)#data should be in dataframe format
colnames(XXXX_RA_data )#verify column names

XXXX_RA_data <- XXXX_RA_data[, sort(colnames(XXXX_RA_data )) ]#sort
colnames(XXXX_RA_data )#double check column names
tail(XXXX_RA_data)
XXXX_RA_data1=XXXX_RA_data # correct format

#get variables with grepl function
idx_01  <- grepl('.*\\.01', colnames(XXXX_RA_data ))
idx_11  <- grepl('.*\\.11', colnames(XXXX_RA_data ))

cbind(colnames(XXXX_RA_data   )[idx_01],colnames(XXXX_RA_data   )[idx_11])#pair   the
##sample columns with cbind and check again matching is correct

#2.run test
#set pair to TRUE for paired testing, exact=FALSE to handle zeros

XXXX_RA_data1$mean_T <- apply(XXXX_RA_data , 1,function(x) mean(as.numeric(x[idx_01])))
XXXX_RA_data1$mean_N <- apply(XXXX_RA_data , 1,function(x) mean(as.numeric(x[idx_11])))

XXXX_RA_data1$sd_T <- apply(XXXX_RA_data , 1,function(x) sd(as.numeric(x[idx_01])))
XXXX_RA_data1$sd_N <- apply(XXXX_RA_data , 1,function(x) sd(as.numeric(x[idx_11])))

XXXX_RA_data1$log2fc <- apply(XXXX_RA_data , 1, function(x) log2(mean(as.numeric(x[idx_01]))) -
log2(mean(as.numeric(x[idx_11]))))

XXXX_RA_data1$pval      <-      apply(XXXX_RA_data      ,      1,      function(x)
wilcox.test(as.numeric(x[idx_01]),  as.numeric(x[idx_11]),  paired  =  T,  exact  =
F)$p.value)

XXXX_RA_data1$fdr<-p.adjust(XXXX_RA_data1$pval, method = "fdr")

write.csv(XXXX_RA_data1, file="Results_XXXX_wilcoxon_paired.csv")

#3. visualize results with log fold change chart (edited from joey711 phyloseq
#tutorial)
#load taxa table as dataframe
XXXX_taxa<-read.csv("XXXX_taxa_table.csv", row.names = 1)
XXXX_taxa<-data.frame(XXXX_taxa)#convert to dataframe

#call results from wilcoxon paired test comparing tumor vs normal
XXXX_RA_data1<-data.frame(XXXX_RA_data1)#call as dataframe
```

```
sigtab<-cbind(XXXX_taxa, XXXX_RA_data1)
alpha <- 0.05
sigtab <- sigtab[(sigtab$pval < alpha), ]
sigtab<- sigtab[order(sigtab$pval),]# sorted by the pvalues fdr values usually all
#1's for our data type


##############################################################################################
#end-
sessionInfo()
```

D.2 Complete list and session info of R-packages used for this project

```
Session info -------------------------------------------------------------- --
------------------

 setting  value
 version  R version 3.5.1 (2018-07-02)
 os       Windows >= 8 x64
 system   x86_64, mingw32
 ui       RStudio
 language (EN)
```

```
 collate  English_United States.1252
 ctype    English_United States.1252
 tz       Pacific/Honolulu
 date     2019-03-14

- Packages -------------------------------------------------------------------
-------------------
 ! package              * version       date       lib source
   abind                  1.4-5         2016-07-21 [1] CRAN (R 3.5.2)
   acepack                1.4.1         2016-10-29 [1] CRAN (R 3.5.1)
   ade4                 * 1.7-13        2018-08-31 [1] CRAN (R 3.5.1)
   annotate               1.60.0        2018-10-30 [1] Bioconductor
   AnnotationDbi          1.44.0        2018-10-30 [1] Bioconductor
   ape                  * 5.2           2018-09-24 [1] CRAN (R 3.5.1)
   assertthat           * 0.2.0         2017-04-11 [1] CRAN (R 3.5.1)
   backports            * 1.1.3         2018-12-14 [1] CRAN (R 3.5.2)
   base64enc              0.1-3         2015-07-28 [1] CRAN (R 3.5.0)
   BiasedUrn            * 1.07          2015-12-28 [1] CRAN (R 3.5.0)
   bindr                * 0.1.1         2018-03-13 [1] CRAN (R 3.5.1)
   bindrcpp             * 0.2.2         2018-03-29 [1] CRAN (R 3.5.1)
   Biobase              * 2.42.0        2018-10-30 [1] Bioconductor
   BiocGenerics         * 0.28.0        2018-10-30 [1] Bioconductor
   BiocInstaller        * 1.32.1        2018-11-01 [1] Bioconductor
   BiocManager          * 1.30.4        2018-11-13 [1] CRAN (R 3.5.1)
   BiocParallel         * 1.16.5        2019-01-03 [1] Bioconductor
   BiocStyle              2.10.0        2018-10-30 [1] Bioconductor
   biomformat             1.10.1        2019-01-04 [1] Bioconductor
   Biostrings             2.50.2        2019-01-03 [1] Bioconductor
   bit                    1.1-14        2018-05-29 [1] CRAN (R 3.5.0)
   bit64                  0.9-7         2017-05-08 [1] CRAN (R 3.5.0)
   bitops               * 1.0-6         2013-08-17 [1] CRAN (R 3.5.0)
   blob                   1.1.1         2018-03-25 [1] CRAN (R 3.5.1)
   boot                 * 1.3-20        2017-08-06 [2] CRAN (R 3.5.1)
   broom                * 0.5.1         2018-12-05 [1] CRAN (R 3.5.1)
   callr                  3.1.1         2018-12-21 [1] CRAN (R 3.5.2)
   car                  * 3.0-2         2018-08-23 [1] CRAN (R 3.5.2)
   carData              * 3.0-2         2018-09-30 [1] CRAN (R 3.5.2)
   caTools                1.17.1.2      2019-03-06 [1] CRAN (R 3.5.2)
   cellranger           * 1.1.0         2016-07-27 [1] CRAN (R 3.5.1)
   checkmate            * 1.9.1         2019-01-15 [1] CRAN (R 3.5.2)
   class                  7.3-15        2019-01-01 [2] CRAN (R 3.5.2)
   cli                  * 1.0.1         2018-09-25 [1] CRAN (R 3.5.1)
   cluster              * 2.0.7-1       2018-04-13 [2] CRAN (R 3.5.1)
   cmprsk                 2.2-7         2014-06-17 [1] CRAN (R 3.5.2)
   coda                   0.19-2        2018-10-08 [1] CRAN (R 3.5.2)
   codetools              0.2-16        2018-12-24 [2] CRAN (R 3.5.2)
   colorspace             1.4-0         2019-01-13 [1] CRAN (R 3.5.2)
   corpcor                1.6.9         2017-04-01 [1] CRAN (R 3.5.0)
   cowplot              * 0.9.4         2019-01-08 [1] CRAN (R 3.5.2)
   crayon               * 1.3.4         2017-09-16 [1] CRAN (R 3.5.1)
   curl                   3.3           2019-01-10 [1] CRAN (R 3.5.2)
   data.table           * 1.12.0        2019-01-13 [1] CRAN (R 3.5.2)
   DBI                    1.0.0         2018-05-02 [1] CRAN (R 3.5.1)
   dbplyr               * 1.3.0         2019-01-09 [1] CRAN (R 3.5.2)
   DelayedArray         * 0.8.0         2018-10-30 [1] Bioconductor
   desc                   1.2.0         2018-05-01 [1] CRAN (R 3.5.1)
   DESeq2               * 1.22.2        2019-01-04 [1] Bioconductor
   devtools             * 2.0.1         2018-10-26 [1] CRAN (R 3.5.1)
   digest               * 0.6.18        2018-10-10 [1] CRAN (R 3.5.1)
   dplyr                * 0.8.0.1       2019-02-15 [1] CRAN (R 3.5.2)
```

```
DT                    0.5         2018-11-05 [1] CRAN (R 3.5.1)
e1071               * 1.7-0.1     2019-01-21 [1] CRAN (R 3.5.2)
edgeR               * 3.24.3      2019-01-02 [1] Bioconductor
epiR                * 0.9-99      2018-11-06 [1] CRAN (R 3.5.1)
evaluate            * 0.13        2019-02-12 [1] CRAN (R 3.5.2)
fansi                 0.4.0       2018-10-05 [1] CRAN (R 3.5.1)
forcats             * 0.4.0       2019-02-17 [1] CRAN (R 3.5.2)
foreach             * 1.4.4       2017-12-12 [1] CRAN (R 3.5.1)
foreign             * 0.8-71      2018-07-20 [2] CRAN (R 3.5.1)
forestplot          * 1.7.2       2017-09-16 [1] CRAN (R 3.5.2)
formatR               1.6         2019-03-05 [1] CRAN (R 3.5.2)
Formula             * 1.2-3       2018-05-03 [1] CRAN (R 3.5.0)
fs                    1.2.6       2018-08-23 [1] CRAN (R 3.5.1)
futile.logger       * 1.4.3       2016-07-10 [1] CRAN (R 3.5.1)
futile.options        1.0.1       2018-04-20 [1] CRAN (R 3.5.0)
gdata                 2.18.0      2017-06-06 [1] CRAN (R 3.5.1)
genefilter          * 1.64.0      2018-10-30 [1] Bioconductor
geneplotter           1.60.0      2018-10-30 [1] Bioconductor
generics            * 0.0.2       2018-11-29 [1] CRAN (R 3.5.1)
GenomeInfoDb        * 1.18.1      2018-11-12 [1] Bioconductor
GenomeInfoDbData      1.2.0       2018-10-11 [1] Bioconductor
GenomicRanges       * 1.34.0      2018-10-30 [1] Bioconductor
GGally              * 1.4.0       2018-05-17 [1] CRAN (R 3.5.2)
ggcorrplot          * 0.1.2       2018-09-11 [1] CRAN (R 3.5.2)
ggfortify           * 0.4.5       2018-05-26 [1] CRAN (R 3.5.2)
ggnetwork           * 0.5.1       2016-03-25 [1] CRAN (R 3.5.2)
ggplot2             * 3.1.0       2018-10-25 [1] CRAN (R 3.5.1)
ggpubr              * 0.2         2018-11-15 [1] CRAN (R 3.5.2)
ggrepel             * 0.8.0       2018-05-09 [1] CRAN (R 3.5.2)
ggsci               * 2.9         2018-05-14 [1] CRAN (R 3.5.2)
ggsignif            * 0.5.0       2019-02-20 [1] CRAN (R 3.5.2)
ggthemes            * 4.1.0       2019-02-19 [1] CRAN (R 3.5.2)
gld                 * 2.4.1       2016-12-05 [1] CRAN (R 3.5.2)
glmnet                2.0-16      2018-04-02 [1] CRAN (R 3.5.1)
glue                * 1.3.0       2018-07-17 [1] CRAN (R 3.5.2)
gmodels               2.18.1      2018-06-25 [1] CRAN (R 3.5.1)
gower               * 0.2.0       2019-03-07 [1] CRAN (R 3.5.2)
gplots                3.0.1.1     2019-01-27 [1] CRAN (R 3.5.2)
gridExtra             2.3         2017-09-09 [1] CRAN (R 3.5.1)
gtable                0.2.0       2016-02-26 [1] CRAN (R 3.5.1)
gtools                3.8.1       2018-06-26 [1] CRAN (R 3.5.0)
GUniFrac            * 1.1         2018-02-12 [1] CRAN (R 3.5.1)
haven                 2.1.0       2019-02-19 [1] CRAN (R 3.5.2)
Hmisc               * 4.2-0       2019-01-26 [1] CRAN (R 3.5.2)
hms                   0.4.2       2018-03-10 [1] CRAN (R 3.5.1)
htmlTable             1.13.1      2019-01-07 [1] CRAN (R 3.5.2)
htmltools           * 0.3.6       2017-04-28 [1] CRAN (R 3.5.1)
htmlwidgets           1.3         2018-09-30 [1] CRAN (R 3.5.1)
httpuv                1.4.5.1     2018-12-18 [1] CRAN (R 3.5.2)
httr                * 1.4.0       2018-12-11 [1] CRAN (R 3.5.2)
igraph              * 1.2.4       2019-02-13 [1] CRAN (R 3.5.2)
infer               * 0.4.0       2018-11-15 [1] CRAN (R 3.5.1)
inline                0.3.15      2018-05-18 [1] CRAN (R 3.5.2)
IRanges             * 2.16.0      2018-10-30 [1] Bioconductor
iterators           * 1.0.10      2018-07-13 [1] CRAN (R 3.5.1)
jsonlite              1.6         2018-12-07 [1] CRAN (R 3.5.1)
KernSmooth            2.23-15     2015-06-29 [2] CRAN (R 3.5.1)
km.ci                 0.5-2       2009-08-30 [1] CRAN (R 3.5.2)
KMsurv                0.1-5       2012-12-03 [1] CRAN (R 3.5.2)
knitr               * 1.22        2019-03-08 [1] CRAN (R 3.5.3)
```

```
labeling             * 0.3          2014-08-23 [1] CRAN (R 3.5.0)
lambda.r               1.2.3        2018-05-17 [1] CRAN (R 3.5.1)
later                  0.8.0        2019-02-11 [1] CRAN (R 3.5.2)
lattice              * 0.20-38      2018-11-04 [1] CRAN (R 3.5.1)
latticeExtra           0.6-28       2016-02-09 [1] CRAN (R 3.5.1)
lazyeval             * 0.2.1        2017-10-29 [1] CRAN (R 3.5.1)
lda                    1.4.2        2015-11-22 [1] CRAN (R 3.5.2)
limma                * 3.38.3       2018-12-02 [1] Bioconductor
lme4                   1.1-21       2019-03-05 [1] CRAN (R 3.5.2)
lmom                   2.7          2019-03-10 [1] CRAN (R 3.5.3)
lmtest               * 0.9-36       2018-04-04 [1] CRAN (R 3.5.1)
locfit                 1.5-9.1      2013-04-20 [1] CRAN (R 3.5.1)
loo                    2.0.0        2018-04-11 [1] CRAN (R 3.5.2)
lubridate            * 1.7.4        2018-04-11 [1] CRAN (R 3.5.1)
magrittr             * 1.5          2014-11-22 [1] CRAN (R 3.5.1)
markdown             * 0.9          2018-12-07 [1] CRAN (R 3.5.2)
MASS                   7.3-51.1     2018-11-01 [1] CRAN (R 3.5.1)
Matrix                 1.2-16       2019-03-08 [2] CRAN (R 3.5.3)
matrixStats          * 0.54.0       2018-07-23 [1] CRAN (R 3.5.1)
memoise                1.1.0        2017-04-21 [1] CRAN (R 3.5.1)
mgcv                 * 1.8-27       2019-02-06 [1] CRAN (R 3.5.2)
microbiome           * 1.4.2        2018-12-01 [1] Bioconductor
mime                   0.6          2018-10-05 [1] CRAN (R 3.5.1)
minqa                  1.2.4        2014-10-09 [1] CRAN (R 3.5.1)
modelr               * 0.1.4        2019-02-18 [1] CRAN (R 3.5.2)
multtest             * 2.38.0       2018-10-30 [1] Bioconductor
munsell              * 0.5.0        2018-06-12 [1] CRAN (R 3.5.1)
network                1.14-377     2019-03-06 [1] CRAN (R 3.5.2)
nlme                 * 3.1-137      2018-04-07 [2] CRAN (R 3.5.1)
nloptr                 1.2.1        2018-10-03 [1] CRAN (R 3.5.1)
nnet                   7.3-12       2016-02-02 [2] CRAN (R 3.5.1)
openxlsx               4.1.0        2018-05-26 [1] CRAN (R 3.5.2)
PathoStat            * 1.8.4        2018-12-02 [1] Bioconductor
permute              * 0.9-5        2019-03-12 [1] CRAN (R 3.5.1)
phyloseq             * 1.26.1       2019-01-04 [1] Bioconductor
pillar               * 1.3.1        2018-12-15 [1] CRAN (R 3.5.2)
pkgbuild               1.0.2        2018-10-16 [1] CRAN (R 3.5.1)
pkgconfig              2.0.2        2018-08-16 [1] CRAN (R 3.5.1)
pkgload                1.0.2        2018-10-29 [1] CRAN (R 3.5.1)
plotly               * 4.8.0        2018-07-20 [1] CRAN (R 3.5.1)
plyr                 * 1.8.4        2016-06-08 [1] CRAN (R 3.5.1)
prettyunits            1.0.2        2015-07-13 [1] CRAN (R 3.5.1)
pROC                 * 1.13.0       2018-09-24 [1] CRAN (R 3.5.2)
processx               3.3.0        2019-03-10 [1] CRAN (R 3.5.3)
promises               1.0.1        2018-04-13 [1] CRAN (R 3.5.1)
ps                     1.3.0        2018-12-21 [1] CRAN (R 3.5.2)
purrr                * 0.3.1        2019-03-03 [1] CRAN (R 3.5.2)
R6                     2.4.0        2019-02-14 [1] CRAN (R 3.5.1)
RColorBrewer           1.1-2        2014-12-07 [1] CRAN (R 3.5.0)
Rcpp                 * 1.0.0        2018-11-07 [1] CRAN (R 3.5.1)
RCurl                * 1.95-4.12    2019-03-04 [1] CRAN (R 3.5.2)
readr                * 1.3.1        2018-12-21 [1] CRAN (R 3.5.2)
readxl               * 1.3.0        2019-02-15 [1] CRAN (R 3.5.2)
remotes                2.0.2        2018-10-30 [1] CRAN (R 3.5.1)
rentrez                1.2.1        2018-03-05 [1] CRAN (R 3.5.1)
reshape              * 0.8.8        2018-10-23 [1] CRAN (R 3.5.2)
reshape2               1.4.3        2017-12-11 [1] CRAN (R 3.5.1)
rhdf5                * 2.26.2       2019-01-02 [1] Bioconductor
Rhdf5lib               1.4.2        2018-12-03 [1] Bioconductor
rio                    0.5.16       2018-11-26 [1] CRAN (R 3.5.2)
```

```
  rlang                 * 0.3.1          2019-01-08 [1] CRAN (R 3.5.2)
  rmarkdown             * 1.11           2018-12-08 [1] CRAN (R 3.5.2)
  ROCR                    1.0-7          2015-03-26 [1] CRAN (R 3.5.1)
  rpart                   4.1-13         2018-02-23 [2] CRAN (R 3.5.1)
  rprojroot               1.3-2          2018-01-03 [1] CRAN (R 3.5.1)
  RSQLite                 2.1.1          2018-05-06 [1] CRAN (R 3.5.1)
  rstan                   2.18.2         2018-11-07 [1] CRAN (R 3.5.2)
  rstudioapi              0.9.0          2019-01-09 [1] CRAN (R 3.5.2)
  RTCGA                 * 1.12.1         2019-01-04 [1] Bioconductor (R 3.5.2)
  RTCGA.clinical        * 20151101.12.0  2018-11-01 [1] Bioconductor (R 3.5.1)
  rvest                   0.3.2          2016-06-17 [1] CRAN (R 3.5.1)
  S4Vectors             * 0.20.1         2018-11-09 [1] Bioconductor
  scales                * 1.0.0          2018-08-09 [1] CRAN (R 3.5.1)
  sessioninfo             1.1.1          2018-11-05 [1] CRAN (R 3.5.1)
  shiny                 * 1.2.0          2018-11-02 [1] CRAN (R 3.5.2)
  shinyjs                 1.0            2018-01-08 [1] CRAN (R 3.5.1)
  sna                     2.4            2016-08-08 [1] CRAN (R 3.5.2)
  StanHeaders             2.18.1         2019-01-28 [1] CRAN (R 3.5.2)
  stargazer             * 5.2.2          2018-05-30 [1] CRAN (R 3.5.2)
  statnet.common          4.2.0          2019-01-08 [1] CRAN (R 3.5.2)
  stringi               * 1.3.1          2019-02-13 [1] CRAN (R 3.5.2)
  stringr               * 1.4.0          2019-02-10 [1] CRAN (R 3.5.2)
  SummarizedExperiment  * 1.12.0         2018-10-30 [1] Bioconductor
  survey                  3.35-1         2019-01-29 [1] CRAN (R 3.5.2)
R survival              * 2.43-3         <NA>       [2] <NA>
  survminer             * 0.4.3          2018-08-04 [1] CRAN (R 3.5.2)
  survMisc                0.5.5          2018-07-05 [1] CRAN (R 3.5.2)
  sva                   * 3.30.1         2019-01-04 [1] Bioconductor
  table1                * 1.1            2018-07-19 [1] CRAN (R 3.5.2)
  tableone              * 0.10.0         2019-02-17 [1] CRAN (R 3.5.2)
  themetagenomics       * 0.1.0          2017-06-06 [1] CRAN (R 3.5.2)
  tibble                * 2.0.1          2019-01-12 [1] CRAN (R 3.5.2)
  tidyr                   0.8.3          2019-03-01 [1] CRAN (R 3.5.2)
  tidyselect              0.2.5          2018-10-11 [1] CRAN (R 3.5.1)
  UpSetR                * 1.3.3          2017-03-21 [1] CRAN (R 3.5.2)
  usethis               * 1.4.0          2018-08-14 [1] CRAN (R 3.5.1)
  utf8                    1.1.4          2018-05-24 [1] CRAN (R 3.5.1)
  vcd                   * 1.4-4          2017-12-06 [1] CRAN (R 3.5.1)
  vegan                 * 2.5-4          2019-02-04 [1] CRAN (R 3.5.2)
  VennDiagram           * 1.6.20         2018-03-28 [1] CRAN (R 3.5.1)
  viridis                 0.5.1          2018-03-29 [1] CRAN (R 3.5.1)
  viridisLite             0.3.0          2018-02-01 [1] CRAN (R 3.5.1)
  webshot                 0.5.1          2018-09-28 [1] CRAN (R 3.5.1)
  withr                 * 2.1.2          2018-03-15 [1] CRAN (R 3.5.1)
  xfun                  * 0.5            2019-02-20 [1] CRAN (R 3.5.1)
  XML                     3.98-1.19      2019-03-06 [1] CRAN (R 3.5.2)
  xml2                  * 1.2.0          2018-01-24 [1] CRAN (R 3.5.1)
  xtable                  1.8-3          2018-08-29 [1] CRAN (R 3.5.1)
  XVector               * 0.22.0         2018-10-30 [1] Bioconductor
  yaml                    2.2.0          2018-07-25 [1] CRAN (R 3.5.1)
  zip                     2.0.0          2019-02-25 [1] CRAN (R 3.5.2)
  zlibbioc                1.28.0         2018-10-30 [1] Bioconductor
  zoo                   * 1.8-4          2018-09-19 [1] CRAN (R 3.5.1)
```

E. Literature Cited, complete list

Arthur, J. C., R. Z. Gharaibeh, M. Muhlbauer, E. Perez-Chanona, J. M. Uronis, J. McCafferty, A. A. Fodor, and C. Jobin. 2014. "Microbial genomic analysis reveals the essential role of inflammation in bacteria-induced colorectal cancer." *Nat Commun* 5:4724. doi: 10.1038/ncomms5724.

Asshauer, K. P., B. Wemheuer, R. Daniel, and P. Meinicke. 2015. "Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data." *Bioinformatics* 31 (17):2882-4. doi: 10.1093/bioinformatics/btv287.

Attie, R., L. T. Chinen, E. M. Yoshioka, M. C. Silva, and V. C. de Lima. 2014. "Acute bacterial infection negatively impacts cancer specific survival of colorectal cancer patients." *World J Gastroenterol* 20 (38):13930-5. doi: 10.3748/wjg.v20.i38.13930.

Banerjee, S., T. Tian, Z. Wei, N. Shih, M. D. Feldman, K. N. Peck, A. M. DeMichele, J. C. Alwine, and E. S. Robertson. 2018. "Distinct Microbial Signatures Associated With Different Breast Cancer Types." *Front Microbiol* 9:951. doi: 10.3389/fmicb.2018.00951.

Banerjee, S., Z. Wei, F. Tan, K. N. Peck, N. Shih, M. Feldman, T. R. Rebbeck, J. C. Alwine, and E. S. Robertson. 2015. "Distinct microbiological signatures associated with triple negative breast cancer." *Sci Rep* 5:15162. doi: 10.1038/srep15162.

Batai, K., A. Bergersen, E. Price, K. Hynes, N. A. Ellis, and B. R. Lee. 2018. "Clinical and Molecular Characteristics and Burden of Kidney Cancer Among Hispanics and Native Americans: Steps Toward Precision Medicine." *Clin Genitourin Cancer*. doi: 10.1016/j.clgc.2018.01.006.

Beuth, J. 2005. "Microorganisms and Cancer." From Friends to Foes, Old Herborn University.

Bhaduri, A., K. Qu, C. S. Lee, A. Ungewickell, and P. A. Khavari. 2012. "Rapid identification of non-human sequences in high-throughput sequencing datasets." *Bioinformatics* 28 (8):1174-5. doi: 10.1093/bioinformatics/bts100.

Bhatt, A. S., S. S. Freeman, A. F. Herrera, C. S. Pedamallu, D. Gevers, F. Duke, J. Jung, M. Michaud, B. J. Walker, S. Young, A. M. Earl, A. D. Kostic, A. I. Ojesina, R. Hasserjian, K. K. Ballen, Y. B. Chen, G. Hobbs, J. H. Antin, R. J. Soiffer, L. R. Baden, W. S. Garrett, J. L. Hornick, F. M. Marty, and M. Meyerson. 2013. "Sequence-based discovery of Bradyrhizobium enterica in cord colitis syndrome." *N Engl J Med* 369 (6):517-28. doi: 10.1056/NEJMoa1211115.

Biarc, J., I. S. Nguyen, A. Pini, F. Gosse, S. Richert, D. Thierse, A. Van Dorsselaer, E. Leize-Wagner, F. Raul, J. P. Klein, and M. Scholler-Guinard. 2004. "Carcinogenic properties of proteins with pro-inflammatory activity from Streptococcus infantarius (formerly S.bovis)." *Carcinogenesis* 25 (8):1477-84. doi: 10.1093/carcin/bgh091.

Bik, E. M., C. D. Long, G. C. Armitage, P. Loomer, J. Emerson, E. F. Mongodin, K. E. Nelson, S. R. Gill, C. M. Fraser-Liggett, and D. A. Relman. 2010. "Bacterial diversity in the oral cavity of 10 healthy individuals." *ISME J* 4 (8):962-74. doi: 10.1038/ismej.2010.30.

Blaser, M. J. 2008. "Understanding microbe-induced cancers." *Cancer Prev Res (Phila)* 1 (1):15-20. doi: 10.1158/1940-6207.CAPR-08-0024.

Bordonaro, M., D. L. Lazarova, and A. C. Sartorelli. 2008. "Butyrate and Wnt signaling: a possible solution to the puzzle of dietary fiber and colon cancer risk?" *Cell Cycle* 7 (9):1178-83. doi: 10.4161/cc.7.9.5818.

Borozan, I., S. Wilson, P. Blanchette, P. Laflamme, S. N. Watt, P. M. Krzyzanowski, F. Sircoulomb, R. Rottapel, P. E. Branton, and V. Ferretti. 2012. "CaPSID: a bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes." *BMC Bioinformatics* 13:206. doi: 10.1186/1471-2105-13-206.

Borozan, I., M. Zapatka, L. Frappier, and V. Ferretti. 2018. "Analysis of Epstein-Barr Virus Genomes and Expression Profiles in Gastric Adenocarcinoma." *J Virol* 92 (2). doi: 10.1128/JVI.01239-17.

Bouvard, V., R. Baan, K. Straif, Y. Grosse, B. Secretan, F. El Ghissassi, L. Benbrahim-Tallaa, N. Guha, C. Freeman, L. Galichet, V. Cogliano, and W. H. O. International Agency for Research on Cancer Monograph Working Group. 2009. "A review of human carcinogens--Part B: biological agents." *Lancet Oncol* 10 (4):321-2.

Brawner, K. M., R. Kumar, C. A. Serrano, T. Ptacek, E. Lefkowitz, C. D. Morrow, D. Zhi, K. R. Kyanam-Kabir-Baig, L. E. Smythies, P. R. Harris, and P. D. Smith. 2017. "Helicobacter pylori infection is associated with an altered gastric microbiota in children." *Mucosal Immunol* 10 (5):1169-1177. doi: 10.1038/mi.2016.131.

Bray, F., A. Jemal, N. Grey, J. Ferlay, and D. Forman. 2012. "Global cancer transitions according to the Human Development Index (2008-2030): a population-based study." *Lancet Oncol* 13 (8):790-801. doi: 10.1016/S1470-2045(12)70211-5.

Picard Tools 2.17.8.

Brooks, A. W., S. Priya, R. Blekhman, and S. R. Bordenstein. 2018. "Gut microbiota diversity across ethnicities in the United States." *PLoS Biol* 16 (12):e2006842. doi: 10.1371/journal.pbio.2006842.

Burnett-Hartman, A. N., P. A. Newcomb, and J. D. Potter. 2008. "Infectious agents and colorectal cancer: a review of Helicobacter pylori, Streptococcus bovis, JC virus, and human papillomavirus." *Cancer Epidemiol Biomarkers Prev* 17 (11):2970-9. doi: 10.1158/1055-9965.EPI-08-0571.

Cancer Genome Atlas, Network. 2012. "Comprehensive molecular portraits of human breast tumours." *Nature* 490 (7418):61-70. doi: 10.1038/nature11412.

Cancer Genome Atlas, Network. 2015. "Comprehensive genomic characterization of head and neck squamous cell carcinomas." *Nature* 517 (7536):576-82. doi: 10.1038/nature14129.

Cancer Genome Atlas Research, Network. 2014. "Comprehensive molecular characterization of gastric adenocarcinoma." *Nature* 513 (7517):202-9. doi: 10.1038/nature13480.

Cancer Genome Atlas Research, Network. 2015. "The Molecular Taxonomy of Primary Prostate Cancer." *Cell* 163 (4):1011-25. doi: 10.1016/j.cell.2015.10.025.

Cantalupo, P. G., J. P. Katz, and J. M. Pipas. 2018. "Viral sequences in human cancer." *Virology* 513:208-216. doi: 10.1016/j.virol.2017.10.017.

Cao, S., M. C. Wendl, M. A. Wyczalkowski, K. Wylie, K. Ye, R. Jayasinghe, M. Xie, S. Wu, B. Niu, R. Grubb, 3rd, K. J. Johnson, H. Gay, K. Chen, J. S. Rader, J. F. Dipersio, F. Chen, and L. Ding. 2016. "Divergent viral presentation among human tumors and adjacent normal tissues." *Sci Rep* 6:28294. doi: 10.1038/srep28294.

Carrick, D. M., M. G. Mehaffey, M. C. Sachs, S. Altekruse, C. Camalier, R. Chuaqui, W. Cozen, B. Das, B. Y. Hernandez, C. J. Lih, C. F. Lynch, H. Makhlouf, P. McGregor, L. M. McShane, J. Phillips Rohan, W. D. Walsh, P. M. Williams, E. M. Gillanders, L. E. Mechanic, and S. D. Schully. 2015. "Robustness of Next Generation Sequencing on Older Formalin-Fixed Paraffin-Embedded Tissue." *PLoS One* 10 (7):e0127353. doi: 10.1371/journal.pone.0127353.

Castellarin, M., R. L. Warren, J. D. Freeman, L. Dreolini, M. Krzywinski, J. Strauss, R. Barnes, P. Watson, E. Allen-Vercoe, R. A. Moore, and R. A. Holt. 2012. "Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma." *Genome Res* 22 (2):299-306. doi: 10.1101/gr.126516.111.

Cavarretta, I., R. Ferrarese, W. Cazzaniga, D. Saita, R. Luciano, E. R. Ceresola, I. Locatelli, L. Visconti, G. Lavorgna, A. Briganti, M. Nebuloni, C. Doglioni, M. Clementi, F. Montorsi, F. Canducci, and A.

Salonia. 2017. "The Microbiome of the Prostate Tumor Microenvironment." *Eur Urol* 72 (4):625-631. doi: 10.1016/j.eururo.2017.03.029.

Caygill, C. P., M. J. Hill, M. Braddick, and J. C. Sharp. 1994. "Cancer mortality in chronic typhoid and paratyphoid carriers." *Lancet* 343 (8889):83-4.

Chan, A. A., M. Bashir, M. N. Rivas, K. Duvall, P. A. Sieling, T. R. Pieber, P. A. Vaishampayan, S. M. Love, and D. J. Lee. 2016. "Characterization of the microbiome of nipple aspirate fluid of breast cancer survivors." *Sci Rep* 6:28061. doi: 10.1038/srep28061.

Chang, A. H., and J. Parsonnet. 2010. "Role of bacteria in oncogenesis." *Clin Microbiol Rev* 23 (4):837-57. doi: 10.1128/CMR.00012-10.

Chen, V. W., C. M. Fenoglio-Preiser, X. C. Wu, R. J. Coates, P. Reynolds, D. L. Wickerham, P. Andrews, C. Hunter, G. Stemmermann, J. S. Jackson, and B. K. Edwards. 1997. "Aggressiveness of colon carcinoma in blacks and whites. National Cancer Institute Black/White Cancer Survival Study Group." *Cancer Epidemiol Biomarkers Prev* 6 (12):1087-93.

Chen, Y., H. Yao, E. J. Thompson, N. M. Tannir, J. N. Weinstein, and X. Su. 2013. "VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue." *Bioinformatics* 29 (2):266-7. doi: 10.1093/bioinformatics/bts665.

Choi, M., U. I. Scholl, W. Ji, T. Liu, I. R. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloglu, S. Ozen, S. Sanjad, C. Nelson-Williams, A. Farhi, S. Mane, and R. P. Lifton. 2009. "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing." *Proc Natl Acad Sci U S A* 106 (45):19096-101. doi: 10.1073/pnas.0910672106.

Cogliano, V. J., R. Baan, K. Straif, Y. Grosse, B. Lauby-Secretan, F. El Ghissassi, V. Bouvard, L. Benbrahim-Tallaa, N. Guha, C. Freeman, L. Galichet, and C. P. Wild. 2011. "Preventable exposures associated with human cancers." *J Natl Cancer Inst* 103 (24):1827-39. doi: 10.1093/jnci/djr483.

Cohen, R. J., B. A. Shannon, J. E. McNeal, T. Shannon, and K. L. Garrett. 2005. "Propionibacterium acnes associated with inflammation in radical prostatectomy specimens: a possible link to cancer evolution?" *J Urol* 173 (6):1969-74. doi: 10.1097/01.ju.0000158161.15277.78.

Contreras, A. V., B. Cocom-Chan, G. Hernandez-Montes, T. Portillo-Bobadilla, and O. Resendis-Antonio. 2016. "Host-Microbiome Interaction and Cancer: Potential Application in Precision Medicine." *Front Physiol* 7:606. doi: 10.3389/fphys.2016.00606.

Cristescu, R., J. Lee, M. Nebozhyn, K. M. Kim, J. C. Ting, S. S. Wong, J. Liu, Y. G. Yue, J. Wang, K. Yu, X. S. Ye, I. G. Do, S. Liu, L. Gong, J. Fu, J. G. Jin, M. G. Choi, T. S. Sohn, J. H. Lee, J. M. Bae, S. T. Kim, S. H. Park, I. Sohn, S. H. Jung, P. Tan, R. Chen, J. Hardwick, W. K. Kang, M. Ayers, D. Hongyue, C. Reinhard, A. Loboda, S. Kim, and A. Aggarwal. 2015. "Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes." *Nat Med* 21 (5):449-56. doi: 10.1038/nm.3850.

Cruz-Munoz, M. E., and E. M. Fuentes-Panana. 2017. "Beta and Gamma Human Herpesviruses: Agonistic and Antagonistic Interactions with the Host Immune System." *Front Microbiol* 8:2521. doi: 10.3389/fmicb.2017.02521.

Culakova, E., R. Thota, M. S. Poniewierski, N. M. Kuderer, A. F. Wogu, D. C. Dale, J. Crawford, and G. H. Lyman. 2014. "Patterns of chemotherapy-associated toxicity and supportive care in US oncology practice: a nationwide prospective cohort study." *Cancer Med* 3 (2):434-44. doi: 10.1002/cam4.200.

Curtis, E., C. Quale, D. Haggstrom, and R. Smith-Bindman. 2008. "Racial and ethnic differences in breast cancer survival: how much is explained by screening, tumor severity, biology, treatment, comorbidities, and demographics?" *Cancer* 112 (1):171-80. doi: 10.1002/cncr.23131.

Daly, G. M., R. M. Leggett, W. Rowe, S. Stubbs, M. Wilkinson, R. H. Ramirez-Gonzalez, M. Caccamo, W. Bernal, and J. L. Heeney. 2015. "Host Subtraction, Filtering and Assembly Validations for Novel Viral Discovery Using Next Generation Sequencing Data." *PLoS One* 10 (6):e0129059. doi: 10.1371/journal.pone.0129059.

Danai, P. A., M. Moss, D. M. Mannino, and G. S. Martin. 2006. "The epidemiology of sepsis in patients with malignancy." *Chest* 129 (6):1432-40. doi: 10.1378/chest.129.6.1432.

Daraei, P., and C. E. Moore. 2015. "Racial Disparity Among the Head and Neck Cancer Population." *J Cancer Educ* 30 (3):546-51. doi: 10.1007/s13187-014-0753-4.

De Flora, S., and P. Bonanni. 2011. "The prevention of infection-associated cancers." *Carcinogenesis* 32 (6):787-95. doi: 10.1093/carcin/bgr054.

De Flora, S., and S. La Maestra. 2015. "Epidemiology of cancers of infectious origin and prevention strategies." *J Prev Med Hyg* 56 (1):E15-20.

de Martel, C., J. Ferlay, S. Franceschi, J. Vignat, F. Bray, D. Forman, and M. Plummer. 2012. "Global burden of cancers attributable to infections in 2008: a review and synthetic analysis." *Lancet Oncol* 13 (6):607-15. doi: 10.1016/S1470-2045(12)70137-7.

de Martel, C., A. E. Llosa, S. M. Farr, G. D. Friedman, J. H. Vogelman, N. Orentreich, D. A. Corley, and J. Parsonnet. 2005. "Helicobacter pylori infection and the risk of development of esophageal adenocarcinoma." *J Infect Dis* 191 (5):761-7. doi: 10.1086/427659.

De Paoli, P., and A. Carbone. 2013. "Carcinogenic viruses and solid cancers without sufficient evidence of causal association." *Int J Cancer* 133 (7):1517-29. doi: 10.1002/ijc.27995.

Deocaris, C. C., N. Widodo, T. Ishii, S. C. Kaul, and R. Wadhwa. 2007. "Functional significance of minor structural and expression changes in stress chaperone mortalin." *Ann N Y Acad Sci* 1119:165-75. doi: 10.1196/annals.1404.007.

DeSantis, C. E., R. L. Siegel, A. G. Sauer, K. D. Miller, S. A. Fedewa, K. I. Alcaraz, and A. Jemal. 2016. "Cancer statistics for African Americans, 2016: Progress and opportunities in reducing racial disparities." *CA Cancer J Clin* 66 (4):290-308. doi: 10.3322/caac.21340.

Deshmukh, S. K., S. Azim, A. Ahmad, H. Zubair, N. Tyagi, S. K. Srivastava, A. Bhardwaj, S. Singh, R. P. Rocconi, and A. P. Singh. 2017. "Biological basis of cancer health disparities: resources and challenges for research." *Am J Cancer Res* 7 (1):1-12.

Deshmukh, S. K., S. K. Srivastava, A. Bhardwaj, A. P. Singh, N. Tyagi, S. Marimuthu, D. L. Dyess, V. Dal Zotto, J. E. Carter, and S. Singh. 2015. "Resistin and interleukin-6 exhibit racially-disparate expression in breast cancer patients, display molecular association and promote growth and aggressiveness of tumor cells through STAT3 activation." *Oncotarget* 6 (13):11231-41. doi: 10.18632/oncotarget.3591.

Dethlefsen, L., M. McFall-Ngai, and D. A. Relman. 2007. "An ecological and evolutionary perspective on human-microbe mutualism and disease." *Nature* 449 (7164):811-8. doi: 10.1038/nature06245.

Elinav, E., R. Nowarski, C. A. Thaiss, B. Hu, C. Jin, and R. A. Flavell. 2013. "Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms." *Nat Rev Cancer* 13 (11):759-71. doi: 10.1038/nrc3611.

Farhana, L., F. Antaki, F. Murshed, H. Mahmud, S. L. Judd, P. Nangia-Makker, E. Levi, Y. Yu, and A. P. Majumdar. 2018. "Gut microbiome profiling and colorectal cancer in African Americans and Caucasian Americans." *World J Gastrointest Pathophysiol* 9 (2):47-58. doi: 10.4291/wjgp.v9.i2.47.

Fosso, B., M. Santamaria, M. D'Antonio, D. Lovero, G. Corrado, E. Vizza, N. Passaro, A. R. Garbuglia, M. R. Capobianchi, M. Crescenzi, G. Valiente, and G. Pesole. 2017. "MetaShot: an accurate workflow

for taxon classification of host-associated microbiome from shotgun metagenomic data." *Bioinformatics* 33 (11):1730-1732. doi: 10.1093/bioinformatics/btx036.

Gagnaire, A., B. Nadel, D. Raoult, J. Neefjes, and J. P. Gorvel. 2017. "Collateral damage: insights into bacterial mechanisms that predispose host cells to cancer." *Nat Rev Microbiol* 15 (2):109-128. doi: 10.1038/nrmicro.2016.171.

Garcia-Castillo, V., E. Sanhueza, E. McNerney, S. A. Onate, and A. Garcia. 2016. "Microbiota dysbiosis: a new piece in the understanding of the carcinogenesis puzzle." *J Med Microbiol* 65 (12):1347-1362. doi: 10.1099/jmm.0.000371.

Global Burden of Disease Cancer, Collaboration, C. Fitzmaurice, D. Dicker, A. Pain, H. Hamavid, M. Moradi-Lakeh, M. F. MacIntyre, C. Allen, G. Hansen, R. Woodbrook, C. Wolfe, R. R. Hamadeh, A. Moore, A. Werdecker, B. D. Gessner, B. Te Ao, B. McMahon, C. Karimkhani, C. Yu, G. S. Cooke, D. C. Schwebel, D. O. Carpenter, D. M. Pereira, D. Nash, D. S. Kazi, D. De Leo, D. Plass, K. N. Ukwaja, G. D. Thurston, K. Yun Jin, E. P. Simard, E. Mills, E. K. Park, F. Catala-Lopez, G. deVeber, C. Gotay, G. Khan, H. D. Hosgood, 3rd, I. S. Santos, J. L. Leasher, J. Singh, J. Leigh, J. B. Jonas, J. Sanabria, J. Beardsley, K. H. Jacobsen, K. Takahashi, R. C. Franklin, L. Ronfani, M. Montico, L. Naldi, M. Tonelli, J. Geleijnse, M. Petzold, M. G. Shrime, M. Younis, N. Yonemoto, N. Breitborde, P. Yip, F. Pourmalek, P. A. Lotufo, A. Esteghamati, G. J. Hankey, R. Ali, R. Lunevicius, R. Malekzadeh, R. Dellavalle, R. Weintraub, R. Lucas, R. Hay, D. Rojas-Rueda, R. Westerman, S. G. Sepanlou, S. Nolte, S. Patten, S. Weichenthal, S. F. Abera, S. M. Fereshtehnejad, I. Shiue, T. Driscoll, T. Vasankari, U. Alsharif, V. Rahimi-Movaghar, V. V. Vlassov, W. S. Marcenes, W. Mekonnen, Y. A. Melaku, Y. Yano, A. Artaman, I. Campos, J. MacLachlan, U. Mueller, D. Kim, M. Trillini, B. Eshrati, H. C. Williams, K. Shibuya, R. Dandona, K. Murthy, B. Cowie, A. T. Amare, C. A. Antonio, C. Castaneda-Orjuela, C. H. van Gool, F. Violante, I. H. Oh, K. Deribe, K. Soreide, L. Knibbs, M. Kereselidze, M. Green, R. Cardenas, N. Roy, T. Tillmann, Y. Li, H. Krueger, L. Monasta, S. Dey, S. Sheikhbahaei, N. Hafezi-Nejad, G. A. Kumar, C. T. Sreeramareddy, L. Dandona, H. Wang, S. E. Vollset, A. Mokdad, J. A. Salomon, R. Lozano, T. Vos, M. Forouzanfar, A. Lopez, C. Murray, and M. Naghavi. 2015. "The Global Burden of Cancer 2013." *JAMA Oncol* 1 (4):505-27. doi: 10.1001/jamaoncol.2015.0735.

Gold, J. S., S. Bayar, and R. R. Salem. 2004. "Association of Streptococcus bovis bacteremia with colonic neoplasia and extracolonic malignancy." *Arch Surg* 139 (7):760-5. doi: 10.1001/archsurg.139.7.760.

Golombos, D. M., A. Ayangbesan, P. O'Malley, P. Lewicki, L. Barlow, C. E. Barbieri, C. Chan, C. DuLong, G. Abu-Ali, C. Huttenhower, and D. S. Scherr. 2018. "The Role of Gut Microbiome in the Pathogenesis of Prostate Cancer: A Prospective, Pilot Study." *Urology* 111:122-128. doi: 10.1016/j.urology.2017.08.039.

Gopalakrishnan, V., B. A. Helmink, C. N. Spencer, A. Reuben, and J. A. Wargo. 2018. "The Influence of the Gut Microbiome on Cancer, Immunity, and Cancer Immunotherapy." *Cancer Cell* 33 (4):570-580. doi: 10.1016/j.ccell.2018.03.015.

Gopalakrishnan, V., C. N. Spencer, L. Nezi, A. Reuben, M. C. Andrews, T. V. Karpinets, P. A. Prieto, D. Vicente, K. Hoffman, S. C. Wei, A. P. Cogdill, L. Zhao, C. W. Hudgens, D. S. Hutchinson, T. Manzo, M. Petaccia de Macedo, T. Cotechini, T. Kumar, W. S. Chen, S. M. Reddy, R. Szczepaniak Sloane, J. Galloway-Pena, H. Jiang, P. L. Chen, E. J. Shpall, K. Rezvani, A. M. Alousi, R. F. Chemaly, S. Shelburne, L. M. Vence, P. C. Okhuysen, V. B. Jensen, A. G. Swennes, F. McAllister, E. Marcelo Riquelme Sanchez, Y. Zhang, E. Le Chatelier, L. Zitvogel, N. Pons, J. L. Austin-Breneman, L. E. Haydu, E. M. Burton, J. M. Gardner, E. Sirmans, J. Hu, A. J. Lazar, T. Tsujikawa, A. Diab, H. Tawbi, I. C. Glitza, W. J. Hwu, S. P. Patel, S. E. Woodman, R. N. Amaria, M. A. Davies, J. E. Gershenwald,

P. Hwu, J. E. Lee, J. Zhang, L. M. Coussens, Z. A. Cooper, P. A. Futreal, C. R. Daniel, N. J. Ajami, J. F. Petrosino, M. T. Tetzlaff, P. Sharma, J. P. Allison, R. R. Jenq, and J. A. Wargo. 2018. "Gut microbiome modulates response to anti-PD-1 immunotherapy in melanoma patients." *Science* 359 (6371):97-103. doi: 10.1126/science.aan4236.

Gourin, C. G., and R. H. Podolsky. 2006. "Racial disparities in patients with head and neck squamous cell carcinoma." *Laryngoscope* 116 (7):1093-106. doi: 10.1097/01.mlg.0000224939.61503.83.

Goyal, S., P. Nangia-Makker, L. Farhana, Y. Yu, and A. P. Majumdar. 2016. "Racial disparity in colorectal cancer: Gut microbiome and cancer stem cells." *World J Stem Cells* 8 (9):279-87. doi: 10.4252/wjsc.v8.i9.279.

Grat, M., K. M. Wronka, M. Krasnodebski, L. Masior, Z. Lewandowski, I. Kosinska, K. Grat, J. Stypulkowski, S. Rejowski, M. Wasilewicz, M. Galecka, P. Szachta, and M. Krawczyk. 2016. "Profile of Gut Microbiota Associated With the Presence of Hepatocellular Cancer in Patients With Liver Cirrhosis." *Transplant Proc* 48 (5):1687-91. doi: 10.1016/j.transproceed.2016.01.077.

Greathouse, K. L., J. R. White, A. J. Vargas, V. V. Bliskovsky, J. A. Beck, N. von Muhlinen, E. C. Polley, E. D. Bowman, M. A. Khan, A. I. Robles, T. Cooks, B. M. Ryan, N. Padgett, A. H. Dzutsev, G. Trinchieri, M. A. Pineda, S. Bilke, P. S. Meltzer, A. N. Hokenstad, T. M. Stickrod, M. R. Walther-Antonio, J. P. Earl, J. C. Mell, J. E. Krol, S. V. Balashov, A. S. Bhat, G. D. Ehrlich, A. Valm, C. Deming, S. Conlan, J. Oh, J. A. Segre, and C. C. Harris. 2018. "Interaction between the microbiome and TP53 in human lung cancer." *Genome Biol* 19 (1):123. doi: 10.1186/s13059-018-1501-6.

Gupta, V. K., S. Paul, and C. Dutta. 2017. "Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity." *Front Microbiol* 8:1162. doi: 10.3389/fmicb.2017.01162.

Ha, J., M. Yan, M. Aguilar, T. Bhuket, M. M. Tana, B. Liu, R. G. Gish, and R. J. Wong. 2016. "Race/ethnicity-specific disparities in cancer incidence, burden of disease, and overall survival among patients with hepatocellular carcinoma in the United States." *Cancer* 122 (16):2512-23. doi: 10.1002/cncr.30103.

Hattori, N., and T. Ushijima. 2016. "Epigenetic impact of infection on carcinogenesis: mechanisms and applications." *Genome Med* 8 (1):10. doi: 10.1186/s13073-016-0267-2.

Hawaii Cancer Center, UHCC. 2016. Hawaii Cancer at a Glance 2009-2013. Honolulu, HI: Hawaii Tumor Registry, University of Hawaii Cancer Center.

Hayashi, K., M. Zhao, K. Yamauchi, N. Yamamoto, H. Tsuchiya, K. Tomita, and R. M. Hoffman. 2009. "Cancer metastasis directly eradicated by targeted therapy with a modified Salmonella typhimurium." *J Cell Biochem* 106 (6):992-8. doi: 10.1002/jcb.22078.

Heath, E. I., F. Lynce, J. Xiu, A. Ellerbrock, S. K. Reddy, E. Obeid, S. V. Liu, A. Bollig-Fischer, D. Separovic, and A. Vanderwalde. 2018. "Racial Disparities in the Molecular Landscape of Cancer." *Anticancer Res* 38 (4):2235-2240. doi: 10.21873/anticanres.12466.

Henao-Mejia, J., E. Elinav, C. A. Thaiss, and R. A. Flavell. 2013. "The intestinal microbiota in chronic liver disease." *Adv Immunol* 117:73-97. doi: 10.1016/B978-0-12-410524-9.00003-7.

Hernandez, B. Y., and M. T. Goodman. 2004. "Ethnic disparities in colorectal cancer incidence and mortality in Hawaii." *Hawaii Med J* 63 (2):54-6.

Hernandez, B. Y., M. T. Goodman, C. F. Lynch, W. Cozen, E. R. Unger, M. Steinau, T. Thompson, M. S. Saber, S. F. Altekruse, C. Lyu, M. Saraiya, and H. P. V. Typing of Cancer Workgroup. 2014. "Human papillomavirus prevalence in invasive laryngeal cancer in the United States." *PLoS One* 9 (12):e115931. doi: 10.1371/journal.pone.0115931.

Hernandez, B. Y., X. Zhu, M. T. Goodman, R. Gatewood, P. Mendiola, K. Quinata, and Y. C. Paulino. 2017. "Betel nut chewing, oral premalignant lesions, and the oral microbiome." *PLoS One* 12 (2):e0172196. doi: 10.1371/journal.pone.0172196.

Hester, C. M., V. R. Jala, M. G. Langille, S. Umar, K. A. Greiner, and B. Haribabu. 2015. "Fecal microbes, short chain fatty acids, and colorectal cancer across racial/ethnic groups." *World J Gastroenterol* 21 (9):2759-69. doi: 10.3748/wjg.v21.i9.2759.

Hoffman, K. L., D. S. Hutchinson, J. Fowler, D. P. Smith, N. J. Ajami, H. Zhao, P. Scheet, W. H. Chow, J. F. Petrosino, and C. R. Daniel. 2018. "Oral microbiota reveals signs of acculturation in Mexican American women." *PLoS One* 13 (4):e0194100. doi: 10.1371/journal.pone.0194100.

Hong, C., S. Manimaran, Y. Shen, J. F. Perez-Rogers, A. L. Byrd, E. Castro-Nallar, K. A. Crandall, and W. E. Johnson. 2014. "PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples." *Microbiome* 2:33. doi: 10.1186/2049-2618-2-33.

Hong, S. N., S. M. Lee, J. H. Kim, T. Y. Lee, J. H. Kim, W. H. Choe, S. Y. Lee, Y. K. Cheon, I. K. Sung, H. S. Park, and C. S. Shim. 2012. "Helicobacter pylori infection increases the risk of colorectal adenomas: cross-sectional study and meta-analysis." *Dig Dis Sci* 57 (8):2184-94. doi: 10.1007/s10620-012-2245-x.

Hoption Cann, S. A., J. P. van Netten, and C. van Netten. 2006. "Acute infections as a means of cancer prevention: opposing effects to chronic infections?" *Cancer Detect Prev* 30 (1):83-93. doi: 10.1016/j.cdp.2005.11.001.

Hornef, M. 2015. "Pathogens, Commensal Symbionts, and Pathobionts: Discovery and Functional Effects on the Host." *ILAR J* 56 (2):159-62. doi: 10.1093/ilar/ilv007.

Howlader, N., S. F. Altekruse, C. I. Li, V. W. Chen, C. A. Clarke, L. A. Ries, and K. A. Cronin. 2014. "US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status." *J Natl Cancer Inst* 106 (5). doi: 10.1093/jnci/dju055.

Hu, H., F. T. Odedina, R. R. Reams, C. T. Lissaker, and X. Xu. 2015. "Racial Differences in Age-Related Variations of Testosterone Levels Among US Males: Potential Implications for Prostate Cancer and Personalized Medication." *J Racial Ethn Health Disparities* 2 (1):69-76. doi: 10.1007/s40615-014-0049-8.

Human Microbiome Project, Consortium. 2012. "Structure, function and diversity of the healthy human microbiome." *Nature* 486 (7402):207-14. doi: 10.1038/nature11234.

Huo, Q., N. Zhang, and Q. Yang. 2012. "Epstein-Barr virus infection and sporadic breast cancer risk: a meta-analysis." *PLoS One* 7 (2):e31656. doi: 10.1371/journal.pone.0031656.

IARC, International Agency for Research in Cancer. 2012. Biological agents. Volume 100 B. A review of human carcinogens. In *IARC Monogr Eval Carcinog Risks Hum*.

Iida, N., A. Dzutsev, C. A. Stewart, L. Smith, N. Bouladoux, R. A. Weingarten, D. A. Molina, R. Salcedo, T. Back, S. Cramer, R. M. Dai, H. Kiu, M. Cardone, S. Naik, A. K. Patri, E. Wang, F. M. Marincola, K. M. Frank, Y. Belkaid, G. Trinchieri, and R. S. Goldszmid. 2013. "Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment." *Science* 342 (6161):967-70. doi: 10.1126/science.1240527.

Iqbal, J., O. Ginsburg, P. A. Rochon, P. Sun, and S. A. Narod. 2015. "Differences in breast cancer stage at diagnosis and cancer-specific survival by race and ethnicity in the United States." *JAMA* 313 (2):165-73. doi: 10.1001/jama.2014.17322.

Isakov, O., S. Modai, and N. Shomron. 2011. "Pathogen detection using short-RNA deep sequencing subtraction and assembly." *Bioinformatics* 27 (15):2027-30. doi: 10.1093/bioinformatics/btr349.

Iyer, P., S. G. Barreto, B. Sahoo, P. Chandrani, M. R. Ramadwar, S. V. Shrikhande, and A. Dutt. 2016. "Non-typhoidal Salmonella DNA traces in gallbladder cancer." *Infect Agent Cancer* 11:12. doi: 10.1186/s13027-016-0057-x.

Jeljaszewicz, J., Gerhard Pulverer, and W. Roszkowski. 1982. *Bacteria and cancer*. London ; New York: Academic Press.

Jiagge, E., A. S. Jibril, D. Chitale, J. M. Bensenhaver, B. Awuah, M. Hoenerhoff, E. Adjei, M. Bekele, E. Abebe, S. D. Nathanson, K. Gyan, B. Salem, J. Oppong, F. Aitpillah, I. Kyei, E. O. Bonsu, E. Proctor, S. D. Merajver, M. Wicha, A. Stark, and L. A. Newman. 2016. "Comparative Analysis of Breast Cancer Phenotypes in African American, White American, and West Versus East African patients: Correlation Between African Ancestry and Triple-Negative Breast Cancer." *Ann Surg Oncol* 23 (12):3843-3849. doi: 10.1245/s10434-016-5420-z.

Jorgensen, S. F., M. Troseid, M. Kummen, J. A. Anmarkrud, A. E. Michelsen, L. T. Osnes, K. Holm, M. L. Hoivik, A. Rashidi, C. P. Dahl, M. Vesterhus, B. Halvorsen, T. E. Mollnes, R. K. Berge, B. Moum, K. E. Lundin, B. Fevang, T. Ueland, T. H. Karlsen, P. Aukrust, and J. R. Hov. 2016. "Altered gut microbiota profile in common variable immunodeficiency associates with levels of lipopolysaccharide and markers of systemic immune activation." *Mucosal Immunol* 9 (6):1455-1465. doi: 10.1038/mi.2016.18.

Kagawa-Singer, M., A. V. Dadia, M. C. Yu, and A. Surbone. 2010. "Cancer, culture, and health disparities: time to chart a new course?" *CA Cancer J Clin* 60 (1):12-39. doi: 10.3322/caac.20051.

Kaminski, J., M. K. Gibson, E. A. Franzosa, N. Segata, G. Dantas, and C. Huttenhower. 2015. "High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED." *PLoS Comput Biol* 11 (12):e1004557. doi: 10.1371/journal.pcbi.1004557.

Kamiya, T., Y. Watanabe, S. Makino, H. Kano, and N. M. Tsuji. 2016. "Improvement of Intestinal Immune Cell Function by Lactic Acid Bacteria for Dairy Products." *Microorganisms* 5 (1). doi: 10.3390/microorganisms5010001.

Kanehisa, M., S. Goto, S. Kawashima, and A. Nakaya. 2002. "The KEGG databases at GenomeNet." *Nucleic Acids Res* 30 (1):42-6.

Karakas, C., C. Wang, F. Deng, H. Huang, D. Wang, and P. Lee. 2017. "Molecular mechanisms involving prostate cancer racial disparity." *Am J Clin Exp Urol* 5 (3):34-48.

Karanth, S., S. S. Rajan, G. Sharma, J. M. Yamal, and R. O. Morgan. 2018. "Racial-Ethnic Disparities in End-of-Life Care Quality among Lung Cancer Patients: A SEER-Medicare-Based Study." *J Thorac Oncol* 13 (8):1083-1093. doi: 10.1016/j.jtho.2018.04.014.

Khoury, J. D., N. M. Tannir, M. D. Williams, Y. Chen, H. Yao, J. Zhang, E. J. Thompson, Tcga Network, F. Meric-Bernstam, L. J. Medeiros, J. N. Weinstein, and X. Su. 2013. "Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq." *J Virol* 87 (16):8916-26. doi: 10.1128/JVI.00340-13.

Kolonel, L. N., D. Altshuler, and B. E. Henderson. 2004. "The multiethnic cohort study: exploring genes, lifestyle and cancer risk." *Nat Rev Cancer* 4 (7):519-27. doi: 10.1038/nrc1389.

Koshiol, J., A. Wozniak, P. Cook, C. Adaniel, J. Acevedo, L. Azocar, A. W. Hsing, J. C. Roa, M. F. Pasetti, J. F. Miquel, M. M. Levine, C. Ferreccio, and Group Gallbladder Cancer Chile Working. 2016. "Salmonella enterica serovar Typhi and gallbladder cancer: a case-control study and meta-analysis." *Cancer Med* 5 (11):3310-3235. doi: 10.1002/cam4.915.

Kostic, A. D., E. Chun, L. Robertson, J. N. Glickman, C. A. Gallini, M. Michaud, T. E. Clancy, D. C. Chung, P. Lochhead, G. L. Hold, E. M. El-Omar, D. Brenner, C. S. Fuchs, M. Meyerson, and W. S. Garrett. 2013. "Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the

tumor-immune microenvironment." *Cell Host Microbe* 14 (2):207-15. doi: 10.1016/j.chom.2013.07.007.

Kostic, A. D., D. Gevers, C. S. Pedamallu, M. Michaud, F. Duke, A. M. Earl, A. I. Ojesina, J. Jung, A. J. Bass, J. Tabernero, J. Baselga, C. Liu, R. A. Shivdasani, S. Ogino, B. W. Birren, C. Huttenhower, W. S. Garrett, and M. Meyerson. 2012. "Genomic analysis identifies association of Fusobacterium with colorectal carcinoma." *Genome Res* 22 (2):292-8. doi: 10.1101/gr.126573.111.

Kostic, A. D., A. I. Ojesina, C. S. Pedamallu, J. Jung, R. G. Verhaak, G. Getz, and M. Meyerson. 2011. "PathSeq: software to identify or discover microbes by deep sequencing of human tissue." *Nat Biotechnol* 29 (5):393-6. doi: 10.1038/nbt.1868.

Kumar, A., P. L. Thotakura, B. K. Tiwary, and R. Krishna. 2016. "Target identification in Fusobacterium nucleatum by subtractive genomics approach and enrichment analysis of host-pathogen protein-protein interactions." *BMC Microbiol* 16:84. doi: 10.1186/s12866-016-0700-0.

Kumar, S., R. Singh, S. Malik, U. Manne, and M. Mishra. 2018. "Prostate cancer health disparities: An immuno-biological perspective." *Cancer Lett* 414:153-165. doi: 10.1016/j.canlet.2017.11.011.

Kuper, H., H. O. Adami, and D. Trichopoulos. 2000. "Infections as a major preventable cause of human cancer." *J Intern Med* 248 (3):171-83.

Laghi, L., A. E. Randolph, D. P. Chauhan, G. Marra, E. O. Major, J. V. Neel, and C. R. Boland. 1999. "JC virus DNA is present in the mucosa of the human colon and in colorectal cancers." *Proc Natl Acad Sci U S A* 96 (13):7484-9.

Lai, C. H., C. S. Chang, H. H. Liu, Y. S. Tsai, F. M. Hsu, Y. L. Yu, C. K. Lai, L. Gandee, R. C. Pong, H. W. Hsu, L. Yu, D. Saha, and J. T. Hsieh. 2014. "Sensitization of radio-resistant prostate cancer cells with a unique cytolethal distending toxin." *Oncotarget* 5 (14):5523-34. doi: 10.18632/oncotarget.2133.

Lakritz, J. R., T. Poutahidis, S. Mirabal, B. J. Varian, T. Levkovich, Y. M. Ibrahim, J. M. Ward, E. C. Teng, B. Fisher, N. Parry, S. Lesage, N. Alberg, S. Gourishetti, J. G. Fox, Z. Ge, and S. E. Erdman. 2015. "Gut bacteria require neutrophils to promote mammary tumorigenesis." *Oncotarget* 6 (11):9387-96. doi: 10.18632/oncotarget.3328.

Langille, M. G., J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. Vega Thurber, R. Knight, R. G. Beiko, and C. Huttenhower. 2013. "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences." *Nat Biotechnol* 31 (9):814-21. doi: 10.1038/nbt.2676.

Leipzig, J. 2017. "A review of bioinformatic pipeline frameworks." *Brief Bioinform* 18 (3):530-536. doi: 10.1093/bib/bbw020.

Lewis, D. A., R. Brown, J. Williams, P. White, S. K. Jacobson, J. R. Marchesi, and M. J. Drake. 2013. "The human urinary microbiome; bacterial DNA in voided urine of asymptomatic adults." *Front Cell Infect Microbiol* 3:41. doi: 10.3389/fcimb.2013.00041.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Subgroup Genome Project Data Processing. 2009. "The Sequence Alignment/Map format and SAMtools." *Bioinformatics* 25 (16):2078-9. doi: 10.1093/bioinformatics/btp352.

Li, J., E. Mercer, X. Gou, and Y. J. Lu. 2013. "Ethnical disparities of prostate cancer predisposition: genetic polymorphisms in androgen-related genes." *Am J Cancer Res* 3 (2):127-51.

Li, J., H. K. Weir, M. A. Jim, S. M. King, R. Wilson, and V. A. Master. 2014. "Kidney cancer incidence and mortality among American Indians and Alaska Natives in the United States, 1990-2009." *Am J Public Health* 104 Suppl 3:S396-403. doi: 10.2105/AJPH.2013.301616.

Love, M. I., W. Huber, and S. Anders. 2014. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biol* 15 (12):550. doi: 10.1186/s13059-014-0550-8.

Lundstig, A., P. Stattin, K. Persson, K. Sasnauskas, R. P. Viscidi, R. E. Gislefoss, and J. Dillner. 2007. "No excess risk for colorectal cancer among subjects seropositive for the JC polyomavirus." *Int J Cancer* 121 (5):1098-102. doi: 10.1002/ijc.22770.

Luo, C., R. Knight, H. Siljander, M. Knip, R. J. Xavier, and D. Gevers. 2015. "ConStrains identifies microbial strains in metagenomic datasets." *Nat Biotechnol* 33 (10):1045-52. doi: 10.1038/nbt.3319.

Mager, D. L. 2006. "Bacteria and cancer: cause, coincidence or cure? A review." *J Transl Med* 4:14. doi: 10.1186/1479-5876-4-14.

Manimaran S, Bendall M, Diaz SV, Castro E, Faits T and Johnson WE 2017. PathoStat: PathoStat Statistical Microbiome Analysis Package. edited by R package.

Marchesi, J. R., B. E. Dutilh, N. Hall, W. H. Peters, R. Roelofs, A. Boleij, and H. Tjalsma. 2011. "Towards the human colorectal cancer microbiome." *PLoS One* 6 (5):e20447. doi: 10.1371/journal.pone.0020447.

Mazouni, C., F. Fina, S. Romain, L. Ouafik, P. Bonnier, J. M. Brandone, and P. M. Martin. 2011. "Epstein-Barr virus as a marker of biological aggressiveness in breast cancer." *Br J Cancer* 104 (2):332-7. doi: 10.1038/sj.bjc.6606048.

Merchant, S. J., L. Li, and J. Kim. 2014. "Racial and ethnic disparities in gastric cancer outcomes: more important than surgical technique?" *World J Gastroenterol* 20 (33):11546-51. doi: 10.3748/wjg.v20.i33.11546.

Miller, J. W., J. L. Smith, A. B. Ryerson, T. C. Tucker, and C. Allemani. 2017. "Disparities in breast cancer survival in the United States (2001-2009): Findings from the CONCORD-2 study." *Cancer* 123 Suppl 24:5100-5118. doi: 10.1002/cncr.30988.

Monographs, IARC. 2012. Biological agents. Volume 100 B. A review of human carcinogens. In *IARC Monogr Eval Carcinog Risks Hum*.

Moore, P. S., and Y. Chang. 2010. "Why do viruses cause cancer? Highlights of the first century of human tumour virology." *Nat Rev Cancer* 10 (12):878-89. doi: 10.1038/nrc2961.

Moore, W. E., and L. H. Moore. 1995. "Intestinal floras of populations that have a high risk of colon cancer." *Appl Environ Microbiol* 61 (9):3202-7.

Murphy, G., A. Michel, P. R. Taylor, D. Albanes, S. J. Weinstein, J. Virtamo, D. Parisi, K. Snyder, J. Butt, K. A. McGlynn, J. Koshiol, M. Pawlita, G. Y. Lai, C. C. Abnet, S. M. Dawsey, and N. D. Freedman. 2014. "Association of seropositivity to Helicobacter species and biliary tract cancer in the ATBC study." *Hepatology* 60 (6):1963-71. doi: 10.1002/hep.27193.

Naccache, S. N., S. Federman, N. Veeraraghavan, M. Zaharia, D. Lee, E. Samayoa, J. Bouquet, A. L. Greninger, K. C. Luk, B. Enge, D. A. Wadford, S. L. Messenger, G. L. Genrich, K. Pellegrino, G. Grard, E. Leroy, B. S. Schneider, J. N. Fair, M. A. Martinez, P. Isa, J. A. Crump, J. L. DeRisi, T. Sittler, J. Hackett, Jr., S. Miller, and C. Y. Chiu. 2014. "A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples." *Genome Res* 24 (7):1180-92. doi: 10.1101/gr.171934.113.

Naeem, R., M. Rashid, and A. Pain. 2013. "READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation." *Bioinformatics* 29 (3):391-2. doi: 10.1093/bioinformatics/bts684.

Nair, N., T. Kasai, and M. Seno. 2014. "Bacteria: prospective savior in battle against cancer." *Anticancer Res* 34 (11):6289-96.

Nauts, H. C. 1989. "Bacteria and cancer--antagonisms and benefits." *Cancer Surv* 8 (4):713-23.

Nauts, H.C. 1982. "Bacterial products in the treatment of cancer: past, present and future." Bacterial and Cancer, Cologne, Germany, 16-18 March, 1982.

NCI. 2015. "The Biology of Cancer Health Disparities." [web], Last Modified 11/09/2015, accessed 02/05. www.cancer.gov/research/progress/discovery/biology-cancer -health-disparities

Newman, L. A. 2017. "Breast Cancer Disparities: Socioeconomic Factors versus Biology." *Ann Surg Oncol* 24 (10):2869-2875. doi: 10.1245/s10434-017-5977-1.

Newman, L. A., and L. M. Kaljee. 2017. "Health Disparities and Triple-Negative Breast Cancer in African American Women: A Review." *JAMA Surg* 152 (5):485-493. doi: 10.1001/jamasurg.2017.0005.

Noecker, C., C. P. McNally, A. Eng, and E. Borenstein. 2017. "High-resolution characterization of the human microbiome." *Transl Res* 179:7-23. doi: 10.1016/j.trsl.2016.07.012.

Nooij, S., D. Schmitz, H. Vennema, A. Kroneman, and M. P. G. Koopmans. 2018. "Overview of Virus Metagenomic Classification Methods and Their Biological Applications." *Front Microbiol* 9:749. doi: 10.3389/fmicb.2018.00749.

Noto, J. M., and R. M. Peek, Jr. 2017. "The gastric microbiome, its interaction with Helicobacter pylori, and its potential role in the progression to stomach cancer." *PLoS Pathog* 13 (10):e1006573. doi: 10.1371/journal.ppat.1006573.

Pandya, D., M. Mariani, S. He, M. Andreoli, M. Spennato, C. Dowell-Martino, P. Fiedler, and C. Ferlini. 2015. "Epstein-Barr Virus MicroRNA Expression Increases Aggressiveness of Solid Malignancies." *PLoS One* 10 (9):e0136058. doi: 10.1371/journal.pone.0136058.

Parise, C. A., K. R. Bauer, M. M. Brown, and V. Caggiano. 2009. "Breast cancer subtypes as defined by the estrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2) among women with invasive breast cancer in California, 1999-2004." *Breast J* 15 (6):593-602. doi: 10.1111/j.1524-4741.2009.00822.x.

Parise, C. A., and V. Caggiano. 2017. "Risk factors associated with the triple-negative breast cancer subtype within four race/ethnicities." *Breast Cancer Res Treat* 163 (1):151-158. doi: 10.1007/s10549-017-4159-y.

Parker, A. S., J. R. Cerhan, C. F. Lynch, B. C. Leibovich, and K. P. Cantor. 2004. "History of urinary tract infection and risk of renal cell carcinoma." *Am J Epidemiol* 159 (1):42-8.

Parkin, D. M. 2006. "The global health burden of infection-associated cancers in the year 2002." *Int J Cancer* 118 (12):3030-44. doi: 10.1002/ijc.21731.

Parsonnet, J. 1995. "Bacterial infection as a cause of cancer." *Environ Health Perspect* 103 Suppl 8:263-8.

Parsonnet, J., I. M. Samloff, L. M. Nelson, N. Orentreich, J. H. Vogelman, and G. D. Friedman. 1993. "Helicobacter pylori, pepsinogen, and risk for gastric adenocarcinoma." *Cancer Epidemiol Biomarkers Prev* 2 (5):461-6.

Paulos, C. M., C. Wrzesinski, A. Kaiser, C. S. Hinrichs, M. Chieppa, L. Cassard, D. C. Palmer, A. Boni, P. Muranski, Z. Yu, L. Gattinoni, P. A. Antony, S. A. Rosenberg, and N. P. Restifo. 2007. "Microbial translocation augments the function of adoptively transferred self/tumor-specific CD8+ T cells via TLR4 signaling." *J Clin Invest* 117 (8):2197-204. doi: 10.1172/JCI32205.

Pellicano, R., V. Mazzaferro, W. F. Grigioni, M. A. Cutufia, S. Fagoonee, L. Silengo, M. Rizzetto, and A. Ponzetto. 2004. "Helicobacter species sequences in liver samples from patients with and without hepatocellular carcinoma." *World J Gastroenterol* 10 (4):598-601.

Pevsner-Fischer, M., T. Tuganbaev, M. Meijer, S. H. Zhang, Z. R. Zeng, M. H. Chen, and E. Elinav. 2016. "Role of the microbiome in non-gastrointestinal cancers." *World J Clin Oncol* 7 (2):200-13. doi: 10.5306/wjco.v7.i2.200.

Piana, A. F., G. Sotgiu, M. R. Muroni, P. Cossu-Rocca, P. Castiglia, and M. R. De Miglio. 2014. "HPV infection and triple-negative breast cancers: an Italian case-control study." *Virol J* 11:190. doi: 10.1186/s12985-014-0190-3.

Plummer, M., C. de Martel, J. Vignat, J. Ferlay, F. Bray, and S. Franceschi. 2016. "Global burden of cancers attributable to infections in 2012: a synthetic analysis." *Lancet Glob Health* 4 (9):e609-16. doi: 10.1016/S2214-109X(16)30143-7.

Porter, Corey M., Eva Shrestha, Lauren B. Peiffer, and Karen S. Sfanos. 2018. "The microbiome in prostate inflammation and prostate cancer." *Prostate Cancer and Prostatic Diseases*. doi: 10.1038/s41391-018-0041-1.

Ragin, C., J. C. Liu, G. Jones, O. Shoyele, B. Sowunmi, R. Kennett, H. J. Groen, D. Gibbs, E. Blackman, M. Esan, M. S. Brandwein, K. Devarajan, F. Bussu, R. Chernock, C. Y. Chien, M. A. Cohen, E. M. Samir, S. Mikio, G. D'Souza, P. Funchain, C. Eng, S. M. Gollin, A. Hong, Y. S. Jung, M. Kruger, J. Lewis, Jr., P. Morbini, S. Landolfo, M. Ritta, J. Straetmans, K. Szarka, R. Tachezy, F. P. Worden, D. Nelson, S. Gathere, and E. Taioli. 2016. "Prevalence of HPV Infection in Racial-Ethnic Subgroups of Head and Neck Cancer Patients." *Carcinogenesis*. doi: 10.1093/carcin/bgw203.

Rauh-Hain, J. A., A. Melamed, D. Schaps, A. J. Bregar, R. Spencer, J. O. Schorge, L. W. Rice, and M. G. Del Carmen. 2018. "Racial and ethnic disparities over time in the treatment and mortality of women with gynecological malignancies." *Gynecol Oncol* 149 (1):4-11. doi: 10.1016/j.ygyno.2017.12.006.

Relman, D. A. 1998. "Detection and identification of previously unrecognized microbial pathogens." *Emerg Infect Dis* 4 (3):382-9. doi: 10.3201/eid0403.980310.

Reuter, J. A., D. V. Spacek, and M. P. Snyder. 2015. "High-throughput sequencing technologies." *Mol Cell* 58 (4):586-97. doi: 10.1016/j.molcel.2015.05.004.

Riley, D. R., K. B. Sieber, K. M. Robinson, J. R. White, A. Ganesan, S. Nourbakhsh, and J. C. Dunning Hotopp. 2013. "Bacteria-human somatic cell lateral gene transfer is enriched in cancer samples." *PLoS Comput Biol* 9 (6):e1003107. doi: 10.1371/journal.pcbi.1003107.

Rios-Covian, D., P. Ruas-Madiedo, A. Margolles, M. Gueimonde, C. G. de Los Reyes-Gavilan, and N. Salazar. 2016. "Intestinal Short Chain Fatty Acids and their Link with Diet and Human Health." *Front Microbiol* 7:185. doi: 10.3389/fmicb.2016.00185.

Robinson, K. M., J. Crabtree, J. S. Mattick, K. E. Anderson, and J. C. Dunning Hotopp. 2017. "Distinguishing potential bacteria-tumor associations from contamination in a secondary data analysis of public cancer genome sequence data." *Microbiome* 5 (1):9. doi: 10.1186/s40168-016-0224-8.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics* 26 (1):139-40. doi: 10.1093/bioinformatics/btp616.

Rogers, Kara. 2016. "Human Microbiome." Encyclopaedia Britannica, Inc., Last Modified April 07, 2016, accessed October 16, 2017. https://www.britannica.com/science/human-microbiome.

Rolston, K. V. 2017. "Infections in Cancer Patients with Solid Tumors: A Review." *Infect Dis Ther* 6 (1):69-83. doi: 10.1007/s40121-017-0146-1.

Routy, B., E. Le Chatelier, L. Derosa, C. P. M. Duong, M. T. Alou, R. Daillere, A. Fluckiger, M. Messaoudene, C. Rauber, M. P. Roberti, M. Fidelle, C. Flament, V. Poirier-Colame, P. Opolon, C. Klein, K. Iribarren, L. Mondragon, N. Jacquelot, B. Qu, G. Ferrere, C. Clemenson, L. Mezquita, J. R. Masip, C. Naltet, S. Brosseau, C. Kaderbhai, C. Richard, H. Rizvi, F. Levenez, N. Galleron, B. Quinquis, N. Pons, B. Ryffel, V. Minard-Colin, P. Gonin, J. C. Soria, E. Deutsch, Y. Loriot, F. Ghiringhelli, G. Zalcman, F. Goldwasser, B. Escudier, M. D. Hellmann, A. Eggermont, D. Raoult, L. Albiges, G. Kroemer, and L. Zitvogel. 2018. "Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors." *Science* 359 (6371):91-97. doi: 10.1126/science.aan3706.

Salyakina, D., and N. F. Tsinoremas. 2013. "Viral expression associated with gastrointestinal adenocarcinomas in TCGA high-throughput sequencing data." *Hum Genomics* 7:23. doi: 10.1186/1479-7364-7-23.

Scelo, G., P. Li, E. Chanudet, and D. C. Muller. 2017. "Variability of Sex Disparities in Cancer Incidence over 30 Years: The Striking Case of Kidney Cancer." *Eur Urol Focus*. doi: 10.1016/j.euf.2017.01.006.

Schmidt, B. L., J. Kuczynski, A. Bhattacharya, B. Huey, P. M. Corby, E. L. Queiroz, K. Nightingale, A. R. Kerr, M. D. DeLacure, R. Veeramachaneni, A. B. Olshen, D. G. Albertson, and Teh Muy-Teck. 2014. "Changes in abundance of oral microbiota associated with oral cancer." *PLoS One* 9 (6):e98741. doi: 10.1371/journal.pone.0098741.

Schwabe, R. F., and C. Jobin. 2013. "The microbiome and cancer." *Nat Rev Cancer* 13 (11):800-12. doi: 10.1038/nrc3610.

Setiawan, V. W., P. C. Wei, B. Y. Hernandez, S. C. Lu, K. R. Monroe, L. Le Marchand, and J. M. Yuan. 2016. "Disparity in liver cancer incidence and chronic liver disease mortality by nativity in Hispanics: The Multiethnic Cohort." *Cancer* 122 (9):1444-52. doi: 10.1002/cncr.29922.

Sfanos, K. S., J. Sauvageot, H. L. Fedor, J. D. Dick, A. M. De Marzo, and W. B. Isaacs. 2008. "A molecular analysis of prokaryotic and viral DNA sequences in prostate tissue from patients with prostate cancer indicates the presence of multiple and diverse microorganisms." *Prostate* 68 (3):306-20. doi: 10.1002/pros.20680.

Shavers, V. L., and M. L. Brown. 2002. "Racial and ethnic disparities in the receipt of cancer treatment." *J Natl Cancer Inst* 94 (5):334-57.

Shevtsov, M., G. Huile, and G. Multhoff. 2018. "Membrane heat shock protein 70: a theranostic target for cancer therapy." *Philos Trans R Soc Lond B Biol Sci* 373 (1738). doi: 10.1098/rstb.2016.0526.

Shuda, M., H. Feng, H. J. Kwun, S. T. Rosen, O. Gjoerup, P. S. Moore, and Y. Chang. 2008. "T antigen mutations are a human tumor-specific signature for Merkel cell polyomavirus." *Proc Natl Acad Sci U S A* 105 (42):16272-7. doi: 10.1073/pnas.0806526105.

Siegel, R. L., S. A. Fedewa, K. D. Miller, A. Goding-Sauer, P. S. Pinheiro, D. Martinez-Tyson, and A. Jemal. 2015. "Cancer statistics for Hispanics/Latinos, 2015." *CA Cancer J Clin* 65 (6):457-80. doi: 10.3322/caac.21314.

Siegel, R. L., K. D. Miller, and A. Jemal. 2015. "Cancer statistics, 2015." *CA Cancer J Clin* 65 (1):5-29. doi: 10.3322/caac.21254.

Siegel, R. L., K. D. Miller, and A. Jemal. 2016. "Cancer statistics, 2016." *CA Cancer J Clin* 66 (1):7-30. doi: 10.3322/caac.21332.

Siegel, R. L., K. D. Miller, and A. Jemal. 2018. "Cancer statistics, 2018." *CA Cancer J Clin* 68 (1):7-30. doi: 10.3322/caac.21442.

Singh, G. K. 2012. "Rural-urban trends and patterns in cervical cancer mortality, incidence, stage, and survival in the United States, 1950-2008." *J Community Health* 37 (1):217-23. doi: 10.1007/s10900-011-9439-6.

Singh, G. K., and A. Jemal. 2017. "Socioeconomic and Racial/Ethnic Disparities in Cancer Mortality, Incidence, and Survival in the United States, 1950-2014: Over Six Decades of Changing Patterns and Widening Inequalities." *J Environ Public Health* 2017:2819372. doi: 10.1155/2017/2819372.

Singh, G. K., M. Siahpush, and S. F. Altekruse. 2013. "Time trends in liver cancer mortality, incidence, and risk factors by unemployment level and race/ethnicity, United States, 1969-2011." *J Community Health* 38 (5):926-40. doi: 10.1007/s10900-013-9703-z.

Sivan, A., L. Corrales, N. Hubert, J. B. Williams, K. Aquino-Michaels, Z. M. Earley, F. W. Benyamin, Y. M. Lei, B. Jabri, M. L. Alegre, E. B. Chang, and T. F. Gajewski. 2015. "Commensal Bifidobacterium

promotes antitumor immunity and facilitates anti-PD-L1 efficacy." *Science* 350 (6264):1084-9. doi: 10.1126/science.aac4255.

Sobhani, I., J. Tap, F. Roudot-Thoraval, J. P. Roperch, S. Letulle, P. Langella, G. Corthier, J. Tran Van Nhieu, and J. P. Furet. 2011. "Microbial dysbiosis in colorectal cancer (CRC) patients." *PLoS One* 6 (1):e16393. doi: 10.1371/journal.pone.0016393.

Spratt, D. E., T. Chan, L. Waldron, C. Speers, F. Y. Feng, O. O. Ogunwobi, and J. R. Osborne. 2016. "Racial/Ethnic Disparities in Genomic Sequencing." *JAMA Oncol* 2 (8):1070-4. doi: 10.1001/jamaoncol.2016.1854.

Steele, C. B., J. Li, B. Huang, and H. K. Weir. 2017. "Prostate cancer survival in the United States by race and stage (2001-2009): Findings from the CONCORD-2 study." *Cancer* 123 Suppl 24:5160-5177. doi: 10.1002/cncr.31026.

Sturtz, L. A., J. Melley, K. Mamula, C. D. Shriver, and R. E. Ellsworth. 2014. "Outcome disparities in African American women with triple negative breast cancer: a comparison of epidemiological and molecular factors between African American and Caucasian women with triple negative breast cancer." *BMC Cancer* 14:62. doi: 10.1186/1471-2407-14-62.

Sun, J., and I. Kato. 2016. "Gut microbiota, inflammation and colorectal cancer." *Genes Dis* 3 (2):130-143. doi: 10.1016/j.gendis.2016.03.004.

Tae, H., E. Karunasena, J. H. Bavarva, L. J. McIver, and H. R. Garner. 2014. "Large scale comparison of non-human sequences in human sequencing data." *Genomics* 104 (6 Pt B):453-8. doi: 10.1016/j.ygeno.2014.08.009.

Tanaka, N., M. Zhao, L. Tang, A. A. Patel, Q. Xi, H. T. Van, H. Takahashi, A. A. Osman, J. Zhang, J. Wang, J. N. Myers, and G. Zhou. 2018. "Gain-of-function mutant p53 promotes the oncogenic potential of head and neck squamous cell carcinoma cells by targeting the transcription factors FOXO3a and FOXM1." *Oncogene* 37 (10):1279-1292. doi: 10.1038/s41388-017-0032-z.

Tang, K. W., B. Alaei-Mahabadi, T. Samuelsson, M. Lindh, and E. Larsson. 2013. "The landscape of viral expression and host gene fusion and adaptation in human cancer." *Nat Commun* 4:2513. doi: 10.1038/ncomms3513.

Telli, M. L., E. T. Chang, A. W. Kurian, T. H. Keegan, L. A. McClure, D. Lichtensztajn, J. M. Ford, and S. L. Gomez. 2011. "Asian ethnicity and breast cancer subtypes: a study from the California Cancer Registry." *Breast Cancer Res Treat* 127 (2):471-8. doi: 10.1007/s10549-010-1173-8.

Thomas, A. M., E. C. Jesus, A. Lopes, S. Aguiar, Jr., M. D. Begnami, R. M. Rocha, P. A. Carpinetti, A. A. Camargo, C. Hoffmann, H. C. Freitas, I. T. Silva, D. N. Nunes, J. C. Setubal, and E. Dias-Neto. 2016. "Tissue-Associated Bacterial Alterations in Rectal Carcinoma Patients Revealed by 16S rRNA Community Profiling." *Front Cell Infect Microbiol* 6:179. doi: 10.3389/fcimb.2016.00179.

Thompson, K. J., J. N. Ingle, X. Tang, N. Chia, P. R. Jeraldo, M. R. Walther-Antonio, K. K. Kandimalla, S. Johnson, J. Z. Yao, S. C. Harrington, V. J. Suman, L. Wang, R. L. Weinshilboum, J. C. Boughey, J. P. Kocher, H. Nelson, M. P. Goetz, and K. R. Kalari. 2017. "A comprehensive analysis of breast cancer microbiota and host gene expression." *PLoS One* 12 (11):e0188873. doi: 10.1371/journal.pone.0188873.

Torre, L. A., A. M. Sauer, M. S. Chen, Jr., M. Kagawa-Singer, A. Jemal, and R. L. Siegel. 2016. "Cancer statistics for Asian Americans, Native Hawaiians, and Pacific Islanders, 2016: Converging incidence in males and females." *CA Cancer J Clin* 66 (3):182-202. doi: 10.3322/caac.21335.

Ursell, L. K., J. L. Metcalf, L. W. Parfrey, and R. Knight. 2012. "Defining the human microbiome." *Nutr Rev* 70 Suppl 1:S38-44. doi: 10.1111/j.1753-4887.2012.00493.x.

van Dijk, E. L., H. Auger, Y. Jaszczyszyn, and C. Thermes. 2014. "Ten years of next-generation sequencing technology." *Trends Genet* 30 (9):418-26. doi: 10.1016/j.tig.2014.07.001.

van Tong, H., P. J. Brindley, C. G. Meyer, and T. P. Velavan. 2017. "Parasite Infection, Carcinogenesis and Human Malignancy." *EBioMedicine* 15:12-23. doi: 10.1016/j.ebiom.2016.11.034.

Vetizou, M., J. M. Pitt, R. Daillere, P. Lepage, N. Waldschmitt, C. Flament, S. Rusakiewicz, B. Routy, M. P. Roberti, C. P. Duong, V. Poirier-Colame, A. Roux, S. Becharef, S. Formenti, E. Golden, S. Cording, G. Eberl, A. Schlitzer, F. Ginhoux, S. Mani, T. Yamazaki, N. Jacquelot, D. P. Enot, M. Berard, J. Nigou, P. Opolon, A. Eggermont, P. L. Woerther, E. Chachaty, N. Chaput, C. Robert, C. Mateus, G. Kroemer, D. Raoult, I. G. Boneca, F. Carbonnel, M. Chamaillard, and L. Zitvogel. 2015. "Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota." *Science* 350 (6264):1079-84. doi: 10.1126/science.aad1329.

Wallace, T. A., D. N. Martin, and S. Ambs. 2011. "Interactions among genes, tumor biology and the environment in cancer health disparities: examining the evidence on a national and global scale." *Carcinogenesis* 32 (8):1107-21. doi: 10.1093/carcin/bgr066.

Wang, H., P. Funchain, G. Bebek, J. Altemus, H. Zhang, F. Niazi, C. Peterson, W. T. Lee, B. B. Burkey, and C. Eng. 2017. "Microbiomic differences in tumor and paired-normal tissue in head and neck squamous cell carcinomas." *Genome Med* 9 (1):14. doi: 10.1186/s13073-017-0405-5.

Wang, Z., M. Gerstein, and M. Snyder. 2009. "RNA-Seq: a revolutionary tool for transcriptomics." *Nat Rev Genet* 10 (1):57-63. doi: 10.1038/nrg2484.

Warren, R. L., D. J. Freeman, S. Pleasance, P. Watson, R. A. Moore, K. Cochrane, E. Allen-Vercoe, and R. A. Holt. 2013. "Co-occurrence of anaerobic bacteria in colorectal carcinomas." *Microbiome* 1 (1):16. doi: 10.1186/2049-2618-1-16.

Weber, G., J. Shendure, D. M. Tanenbaum, G. M. Church, and M. Meyerson. 2002. "Identification of foreign gene sequences by transcript filtering against the human genome." *Nat Genet* 30 (2):141-2. doi: 10.1038/ng818.

White, M. C., D. K. Espey, J. Swan, C. L. Wiggins, C. Eheman, and J. S. Kaur. 2014. "Disparities in cancer mortality and incidence among American Indians and Alaska Natives in the United States." *Am J Public Health* 104 Suppl 3:S377-87. doi: 10.2105/AJPH.2013.301673.

WHO, World Health Organization. 2018. "Cancer: Key facts." WHO. https://www.who.int/en/news-room/fact-sheets/detail/cancer.

Williams, V. L., S. Awasthi, A. K. Fink, J. M. Pow-Sang, J. Y. Park, T. Gerke, and K. Yamoah. 2018. "African-American men and prostate cancer-specific mortality: a competing risk analysis of a large institutional cohort, 1989-2015." *Cancer Med*. doi: 10.1002/cam4.1451.

Wotherspoon, A. C., C. Ortiz-Hidalgo, M. R. Falzon, and P. G. Isaacson. 1991. "Helicobacter pylori-associated gastritis and primary B-cell gastric lymphoma." *Lancet* 338 (8776):1175-6.

Wu, G. D., J. Chen, C. Hoffmann, K. Bittinger, Y. Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, H. Li, F. D. Bushman, and J. D. Lewis. 2011. "Linking long-term dietary patterns with gut microbial enterotypes." *Science* 334 (6052):105-8. doi: 10.1126/science.1208344.

Xia, L. C., J. A. Cram, T. Chen, J. A. Fuhrman, and F. Sun. 2011. "Accurate genome relative abundance estimation based on shotgun metagenomic reads." *PLoS One* 6 (12):e27992. doi: 10.1371/journal.pone.0027992.

Xie, G., X. Wang, P. Liu, R. Wei, W. Chen, C. Rajani, B. Y. Hernandez, R. Alegado, B. Dong, D. Li, and W. Jia. 2016. "Distinctly altered gut microbiota in the progression of liver disease." *Oncotarget* 7 (15):19355-66. doi: 10.18632/oncotarget.8466.

Xu, J., C. Xiang, C. Zhang, B. Xu, J. Wu, R. Wang, Y. Yang, L. Shi, J. Zhang, and Z. Zhan. 2019. "Microbial biomarkers of common tongue coatings in patients with gastric cancer." *Microb Pathog* 127:97-105. doi: 10.1016/j.micpath.2018.11.051.

Xu, X., X. Chen, H. Hu, A. B. Dailey, and B. D. Taylor. 2015. "Current opinion on the role of testosterone in the development of prostate cancer: a dynamic model." *BMC Cancer* 15:806. doi: 10.1186/s12885-015-1833-5.

Xu, Y., N. Stange-Thomann, G. Weber, R. Bo, S. Dodge, R. G. David, K. Foley, J. Beheshti, N. L. Harris, B. Birren, E. S. Lander, and M. Meyerson. 2003. "Pathogen discovery from human tissue by sequence-based computational subtraction." *Genomics* 81 (3):329-35.

Xuan, C., J. M. Shamonki, A. Chung, M. L. Dinome, M. Chung, P. A. Sieling, and D. J. Lee. 2014. "Microbial dysbiosis is associated with human breast cancer." *PLoS One* 9 (1):e83744. doi: 10.1371/journal.pone.0083744.

Yamoah, K., M. H. Johnson, V. Choeurng, F. A. Faisal, K. Yousefi, Z. Haddad, A. E. Ross, M. Alshalafa, R. Den, P. Lal, M. Feldman, A. P. Dicker, E. A. Klein, E. Davicioni, T. R. Rebbeck, and E. M. Schaeffer. 2015. "Novel Biomarker Signature That May Predict Aggressive Disease in African American Men With Prostate Cancer." *J Clin Oncol* 33 (25):2789-96. doi: 10.1200/JCO.2014.59.8912.

Yow, M. A., S. N. Tabrizi, G. Severi, D. M. Bolton, J. Pedersen, BioResource Australian Prostate Cancer, G. G. Giles, and M. C. Southey. 2017. "Characterisation of microbial communities within aggressive prostate cancer tissues." *Infect Agent Cancer* 12:4. doi: 10.1186/s13027-016-0112-7.

Yu, G., J. Torres, N. Hu, R. Medrano-Guzman, R. Herrera-Goepfert, M. S. Humphrys, L. Wang, C. Wang, T. Ding, J. Ravel, P. R. Taylor, C. C. Abnet, and A. M. Goldstein. 2017. "Molecular Characterization of the Human Stomach Microbiota in Gastric Cancer Patients." *Front Cell Infect Microbiol* 7:302. doi: 10.3389/fcimb.2017.00302.

Yu, H., H. Meng, F. Zhou, X. Ni, S. Shen, and U. N. Das. 2015. "Urinary microbiota in patients with prostate cancer and benign prostatic hyperplasia." *Arch Med Sci* 11 (2):385-94. doi: 10.5114/aoms.2015.50970.

Zahnd, W. E., A. J. Fogleman, and W. D. Jenkins. 2018. "Rural-Urban Disparities in Stage of Diagnosis Among Cancers With Preventive Opportunities." *Am J Prev Med*. doi: 10.1016/j.amepre.2018.01.021.

Zhang, C., K. Cleveland, F. Schnoll-Sussman, B. McClure, M. Bigg, P. Thakkar, N. Schultz, M. A. Shah, and D. Betel. 2015. "Identification of low abundance microbiome in clinical samples using whole genome sequencing." *Genome Biol* 16:265. doi: 10.1186/s13059-015-0821-z.

Zhang, W., A. Edwards, E. K. Flemington, and K. Zhang. 2017. "Racial disparities in patient survival and tumor mutation burden, and the association between tumor mutation burden and cancer incidence rate." *Sci Rep* 7 (1):13639. doi: 10.1038/s41598-017-13091-y.

Zhu, H., Z. Shen, H. Luo, W. Zhang, and X. Zhu. 2016. "Chlamydia Trachomatis Infection-Associated Risk of Cervical Cancer: A Meta-Analysis." *Medicine (Baltimore)* 95 (13):e3077. doi: 10.1097/MD.0000000000003077.

Zitvogel, L., R. Daillere, M. P. Roberti, B. Routy, and G. Kroemer. 2017. "Anticancer effects of the microbiome and its products." *Nat Rev Microbiol* 15 (8):465-478. doi: 10.1038/nrmicro.2017.44.